Hepting et al v. AT&T Corp. et al Doc. 294

ELECTRONIC FRONTIER FOUNDATION
CINDY COHN (145997)
cindy@eff.org
LEE TIEN (148216)
tien@eff.org
KURT OPSAHL (191303)
kurt@eff.org
KEVIN S. BANKSTON (217026)
bankston@eff.org
CORYNNE MCSHERRY (221504)
corynne@eff.org
JAMES S. TYRE (083117)
jstyre@eff.org
454 Shotwell Street
San Francisco, CA 94110
Telephone: 415/436-9333
415/436-9993 (fax)

TRABER & VOORHEES
BERT VOORHEES (137623)
bv@tvlegal.com
THERESA M. TRABER (116305)
tmt@tvlegal.com
128 North Fair Oaks Avenue, Suite 204
Pasadena, CA 91103
Telephone: 626/585-9611
626/ 577-7079 (fax)

LAW OFFICE OF RICHARD R. WIEBE
RICHARD R. WIEBE (121156)
wiebe@pacbell.net
425 California Street, Suite 2025
San Francisco, CA 94104
Telephone: 415/433-3200
415/433-6382 (fax)

Attorneys for Plaintiffs

[Additional counsel appear on signature page.]

UNITED STATES DISTRICT COURT

FOR THE NORTHERN DISTRICT OF CALIFORNIA

TASH HEPTING, GREGORY HICKS, CAROLYN JEWEL and ERIK KNUTZEN, on Behalf of Themselves and All Others Similarly Situated,,

Plaintiffs,

v.

AT&T CORP., et al.,

Defendants.

No. C-06-0672-VRW

CLASS ACTION

EXHIBITS A-K, Q-T, AND V-Y TO DECLARATION OF J. SCOTT MARCUS IN SUPPORT OF PLAINTIFFS' MOTION FOR PRELIMINARY INJUNCTION

Date: June 8, 2006
Courtroom: 6, 17th Floor
Judge: Hon. Vaughn Walker

Dockets.Justia.com

Exhibit A

CV – J. Scott Marcus

Date of birth:	05.12.1949	
Nationality:	U.S.	
Civil status:	Divorced	
Higher education	1968 - 1972	City College of New York (CUNY), studies in Political Science with a concentration in Public Administration. Strong minor (29 credits) in Computer Science. 3.3/4.0 GPA (cum laude).
	1976 – 1980	Columbia University, School of Engineering. Concentration in Computer Engineering. 3.7/4.0 GPA.

Academic qualifications:

B.A. in Political Science (Public Administration) from the City College of New York.

M.A. from the School of Engineering, Columbia University.

Language skills: English, moderate German, some French.

Present position: Senior Consultant for WIK-Consult Gmbh.

Key qualifications: More than thirty years in positions of progressively greater responsibility in industry and government. Experience in policy analysis, engineering, sales, marketing, financial analysis, and consulting. Facile in addressing the engineering, legal, and economic aspects of telecommunications regulation in an integrated manner. A seasoned public speaker with a significant record of publications.

Professional Experience:

7/2005 – Present

WIK-Consult Gmbh

Senior Consultant

Analyzed U.S. and Canadian experience with flexible spectrum management for the German Regulatory Authority, the BNetzA (2005).

Conducted a study of network interconnection in an NGN environment, with an emphasis on developments in the US and UK, for the BNetzA (2005).

Currently serve as WIK's project manager for a study of collective use of spectrum (including licence-exempt commons) for the European Commission (joint with Mott Macdonald, Aegis, Indepen, and IDATE).

Contribute to the organization of WIK workshops and events, including "NGN and Emerging Markets" (December 2005) and "Bill and Keep" (April 2006).

7/2005 – Present

Independent Consultant

Serve as an advisor to senior management at the Jamaican regulatory authority, the OUR. Primary areas of interest are broadband deployment and adoption, ICT development generally, network interconnection, and Internet issues.

Have been commissioned to prepare a report on "Interconnection in an NGN environment" for presentation at an ITU-T workshop in March 2006.

7/2001 – 6/2005
Federal Communications Commission (FCC), Washington, DC, USA

Senior Advisor for Internet Technology

This was a senior staff position, comparable to the Chief Economist or the Chief Technologist. Primary function was to provide technical and policy advice to FCC senior management in regard to regulation, and non-regulation, of the Internet. Served as the Commission's leading expert in Internet matters. Contributed to proceedings related to Voice over IP (VoIP), lawful intercept over the Internet (CALEA), broadband deployment and adoption, and intercarrier compensation. Represented the FCC in various inter-agency fora related to Internet matters.

10/2003 and 2/2004-7/2004

German Marshall Fund of the United States, Brussels, Belgium

Transatlantic Fellow

Was granted leave from the FCC to study the European Union's new regulatory framework for electronic communications. Worked with the European Commission (unit DG INFSO B1) on the public consultation on IP telephony. Wrote several papers and gave numerous talks.

3/1990 – 7/2001

GTE Internetworking (Genuity), Burlington, MA, USA

Chief Technology Officer (CTO)

Primary duties in this role were:

- regulatory and public policy advocacy;
- technology-related publicity and public outreach;
- oversight of Genuity's participation in standards bodies and industry fora;
- oversight of GTE Internetworking's research agenda,

which included all research performed on our behalf by GTE Labs; and

- management of technical functions that benefit from direct sponsorship of a senior executive, including Internet interconnection (peering) policy and network security strategy.

8/1989 – 3/1990

Sun Microsystems, Billerica, MA

Engineering Manager

Headed a team of a dozen senior developers in the creation of new versions of PC/NFS, Sun's network file system platform, for MS-DOS and OS|2. Had matrixed responsibility for SQA and documentation personnel.

1/1985 – 8/1989

Nixdorf Computer AG, Santa Clara, CA and Munich, Germany

Manager, Communication Software Development

Managed a small team of network software developers in a joint project with Amdahl Corporation. Consulted for Nixdorf in Munich, Germany, and managed data communication software developers.

10/1981 – 1/1985

Spartacus Computers, Inc., Burlington, MA

Director, Software Development

A technical founder and in charge of software development for an innovative start-up company. Developed the first commercial TCP/IP Ethernet solution for IBM mainframes under the VM operating system.

4/1981 – 10/1981

Interactive Data Corporation, Waltham, MA

Manager of Capacity Planning

Managed and trained a small staff of computer performance analysts, and was heavily involved in both strategic and tactical hardware and software planning for the data center. Introduced the use of regression-based forecasting methods, queuing theory, discrete event simulation, and new software monitoring tools.

1/1979 – 4/1981

The Analytic Sciences Corp. (TASC), Reading, MA

Manager of Systems Programming

Managed a small group of systems programmers, maintained and enhanced TASC's heavily modified IBM VM/CMS and VS1 operating systems.

4/1974 – 1/1979

SyncSort, Inc., Englewood Cliffs, NJ

Product Manager

Sales and sales management for a systems software product to sort and sequence data.

1/1968 – 4/1974

New York City Government, New York, NY

Consultant / Systems Project Leader Managed a team of systems programmers for the NYC Health Services Administration. Developed applications programs for the Office of the Mayor, in a variety of computer languages. Created innovative computer algorithms for processing of spatially correlated data.

Membership, Activities:

Co-editor for public policy and regulation for *IEEE Communications Magazine*, program committee member for the TPRC conference, former member of IEEE ComSoc Meetings and Conference Board, former Chair of IEEE CNOM. Senior Member of the IEEE. Former trustee of the American Registry of Internet Numbers (ARIN) from 2000 to 2002.

Main Publications and Conference Presentations:

Publications:

“Interconnection in an NGN environment”, forthcoming, commissioned by the ITU-T for presentation at their ITU New Initiatives Workshop on “What Rules for IP-enabled NGNs?”, March 23-24 2005.

With Lorenz Nett, Mark Scanlan, Ulrich Stumpf, Martin Cave and Gerard Pogorel, Towards More Flexible Spectrum Regulation, a WIK study for the German BNetzA. Available at: http://www.bundesnetzagentur.de/media/archive/4745.pdf. Also available in German.

“Voice over IP (VoIP) and Access to Emergency Services: A Comparison between the U.S. and the European Union”, to appear in IEEE Communications Magazine.

“Is the U.S. Dancing to a Different Drummer?”, Communications & Strategies, no. 60, 4th quarter 2005. Available at:

http://www.idate.fr/fic/revue_telech/132/CS60%20MARCUS.pdf

With Justus Haucap, "Why Regulate? Lessons from New Zealand", IEEE Communications Magazine, November 2005, at: http://www.comsoc.org/ci1/Public/2005/nov/ (click on "Regulatory and Policy").

With Douglas C. Sicker, "Layers Revisited", presented at TPRC, September 2005, available at: http://web.si.umich.edu/tprc/papers/2005/492/Layers%20Revisited%20v0.4.pdf.

"Structured Legislation", in preparation.

"Procompetitive Regulation and Broadband Adoption in Europe", in preparation.

"Beyond Layers", to appear in the *Journal on Telecommunications and High Technology Law*, 2006.

"Broadband Adoption in Europe", *IEEE Communications Magazine*, April 2005, available at: http://www.comsoc.org/ci1/Public/2005/apr/ (click on "Regulatory and Policy" on the left margin of the page).

"The Challenge of Telephone Call Termination Fees", *Enterprise Europe*, January 2005. Available at: http://www.european-enterprise.org/public/docs/EEJ.pdf.

"Universal Service in a Changing World", *IEEE Communications Magazine*, January 2005, available at: http://www.comsoc.org/ci1/Public/2005/jan/ (click on "Regulatory and Policy" on the left margin of the page).

"Europe's New Regulatory Framework for Electronic Communications in Action", presented at the 4th ZEW Conference on the Economics of Information and Communication Technologies, Mannheim, Germany, July 2004. Available at: ftp://ftp.zew.de/pub/zew-docs/div/IKT04/Paper_Marcus_Invited.pdf.

"Call Termination Fees: The U.S. in global perspective", presented at the 4th ZEW Conference on the Economics of Information and Communication Technologies, Mannheim, Germany, July 2004. Available at: ftp://ftp.zew.de/pub/zew-docs/div/IKT04/Paper_Marcus_Parallel_Session.pdf.

"Evolving Core Capabilities of the Internet", *Journal on Telecommunications and High Technology Law*, 2004.

Federal Communications Commission (FCC) Office of Strategic Planning and Policy Analysis (OSP) Working Paper 36, "The Potential Relevance to the United States of the European Union's Newly Adopted Regulatory Framework for Telecommunications," July 2002, available at http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-224213A2.pdf. The article and derivative works also appear in: *Rethinking Rights and Regulations: Institutional Responses to New Communications Technologies*, Ed. Lorrie Faith Cranor and Steven S.

Wildman, MIT Press, 2003; in the *Journal on Telecommunications and High Technology Law* 111 (2003); and in the *2004 Annual Review* of the European Competitive Telecommunications Association (ECTA).

With Jean-Jacques Laffont, Patrick Rey, and Jean Tirole, IDE-I, Toulouse, "Internet interconnection and the off-net-cost pricing principle", *RAND Journal of Economics*, Vol. 34, No. 2, Summer 2003. An earlier version of the paper appeared as "Internet Peering", *American Economics Review*, Volume 91, Number 2, May 2001.

Designing Wide Area Networks and Internetworks: A Practical Guide, Addison Wesley, 1999.

"Internet Hardware and Software", *Proc. IEEE Electro '96*, 1996.

"Icaros, Alice, and the OSF DME", *Proc. of ISINM '95*.An earlier version appeared in *Proc. Fifth IFIP International Workshop on Distributed Systems: Operations and Management* (DSOM '94), October 1994.

"OSI Network Integration:Seamless, or Seamy?", *Proc. of the International Space Year (ISY) Conference on Earth and Space Science Information Systems* (ESSIS), February 1992.

With Lent, R., "An Implementation Architecture for UNIX™ STREAMS-Based Communications Software", Nixdorf technical report.

"Why an SNA PU 5?", Nixdorf technical report.

"KNET:A TCP/IP for the IBM/370", *Proc. IEEE Infocom '87*, March 1987.

With Mower, J., and White, C., "Designing an Ethernet Interface for the System/370", *Proc. IEEE CompCon*, September 1982.

With Mower, J., Malnati, P., and White, C., "System/370 Ethernet Interface Architecture", Spartacus Technical Report 820601, June 1982.

"Performance Analysis of Local Area Networks", *Proc. SHARE 60*, 1982.

"Analysis of DASD Performance", *Proc. SHARE 57*, 1980.

Presentations:

Is the U.S. Dancing to a Different Drummer?, IDATE Transatlantic Telecommunications Dialogue, Montpellier, France, November 22, 2005.

VoIP and European Regulation, U.S. Department of State, Washington, DC, USA, June 3, 2005.

Beyond Layers, Silicon Flatirons Conference, Boulder, Colorado, USA, February, 2005.

Internet Peering, World Bank, Washington, DC, USA, February, 2005.

VoIP: A Massive Paradigm Shift, IIR VoIP World Congress, November 15, 2004.

U.S. Perspectives on European Regulation of Electronic Communications, European Internet Foundation, Brussels, Belgium, November 10, 2004.

Economics of Network Design, FCC internal classes, Washington, DC, USA, October 26 and November 2, 2004.

Evolving the Core: Deployment Challenges and the Internet, North American Network Operators' Group (NANOG), Washington, DC, USA, October 19, 2004.

Europe's New Regulatory Framework for Electronic Communications in Action, 4th ZEW Conference on the Economics of Information and Communication Technologies, Mannheim, Germany, July 2004.

Call Termination Fees: The U.S. in global perspective, presented at the 4th ZEW Conference on the Economics of Information and Communication Technologies, Mannheim, Germany, July 2004.

FTTH: A Transatlantic Regulatory Perspective, FTTH Council, Brussels, Belgium, June 2004.

IP Telephony: Regulatory Challenges, VON Europe, London, UK, June 8, 2004. Updated version of the same talk, Georgetown University, October 7, 2004.

Broadband Policy US and EU, ITU All Star Network Access Symposium, Geneva, Switzerland, June 4, 2004.

Regulation in a Converging World: A Comparison between the EU and the US, ETNO Senior Executive Conference, Warsaw, Poland, May 14, 2004.

Europe's New Regulatory Framework for Electronic Communications: A U.S. Perspective, WIK conference on European regulatory framework, Berlin, Germany, November, 2003. Same talk a few days later, British Institute of Comparative and International Law (BIICL), London, England.

CALEA, the Internet and the FCC, Telestrategies: Intelligence Support for Lawful Interception and Internet Surveillance, Washington, DC, USA, November 13, 2003.

Facilities-Based Aspects of Broadband Deployment in the U.S., Vision in Business: Telecommunications Regulation and Competition Law, Brussels, Belgium, October 23, 2003.

Will Internet Telephony Bring About a Revolution in Telecom Policy?, CATO Institute, September 9, 2003.

Internet Access for the Caribbean, First Jamaica Internet Forum, Ocho Rios, Jamaica, February, 2003.

Global Traffic Exchange among Internet Service Providers, OECD, Berlin, Germany, June 7, 2001.

Exhibit B

The New York Times Archive

NYTimes Go to a Section Go Welcome, barakweinstein - Member Center - Log out

SEARCH NYT Since 1981 Search TimesSelect FREE 14-DAY TRIAL!

Tip for TimesSelect subscribers: Want to easily save this page? Use Times File by simply clicking on the Save Article icon in the Article Tools box below.

NATIONAL DESK

DOMESTIC SURVEILLANCE: THE PROGRAM; SPY AGENCY MINED VAST DATA TROVE, OFFICIALS REPORT

By ERIC LICHTBLAU AND JAMES RISEN (NYT) 1288 words
Published: December 24, 2005

WASHINGTON, Dec. 23 - The National Security Agency has traced and analyzed large volumes of telephone and Internet communications flowing into and out of the United States as part of the eavesdropping program that President Bush approved after the Sept. 11, 2001, attacks to hunt for evidence of terrorist activity, according to current and former government officials.

The volume of information harvested from telecommunication data and voice networks, without court-approved warrants, is much larger than the White House has acknowledged, the officials said. It was collected by tapping directly into some of the American telecommunication system's main arteries, they said.

As part of the program approved by President Bush for domestic surveillance without warrants, the N.S.A. has gained the cooperation of American telecommunications companies to obtain backdoor access to streams of domestic and international communications, the officials said.

The government's collection and analysis of phone and Internet traffic have raised questions among some law enforcement and judicial officials familiar with the program. One issue of concern to the Foreign Intelligence Surveillance Court, which has reviewed some separate warrant applications growing out of the N.S.A.'s surveillance program, is whether the court has legal authority over calls outside the United States that happen to pass through American-based telephonic "switches," according to officials familiar with the matter.

"There was a lot of discussion about the switches" in conversations with the court, a Justice Department official said, referring to the gateways through which much of the communications traffic flows. "You're talking about access to such a vast amount of communications, and the question was, How do you minimize something that's on a switch that's carrying such large volumes of traffic? The court was very, very concerned about that."

Since the disclosure last week of the N.S.A.'s domestic surveillance program, President Bush and his senior aides have stressed that his executive order allowing eavesdropping without warrants was limited to the monitoring of international phone and e-mail communications involving people with known links to Al Qaeda.

What has not been publicly acknowledged is that N.S.A. technicians, besides actually eavesdropping on specific conversations, have combed through large volumes of phone and Internet traffic in search of patterns that might point to terrorism suspects. Some officials describe the program as a large data-mining operation.

The current and former government officials who discussed the program were granted anonymity because it remains classified.

Bush administration officials declined to comment on Friday on the technical aspects of the operation and the N.S.A.'s use of broad searches to look for clues on terrorists. Because the program is highly classified, many details of how the N.S.A. is conducting it remain unknown, and members of Congress who have pressed for a full Congressional inquiry say they are eager to learn more about the program's operational details, as well as its legality.

Officials in the government and the telecommunications industry who have knowledge of parts of the program say the N.S.A. has sought to analyze communications patterns to glean clues from details like who is calling whom, how long a phone call lasts and what time of day it is made, and the origins and destinations of phone calls and e-mail messages. Calls to and from Afghanistan, for instance, are known to have been of particular interest to the N.S.A. since the Sept. 11 attacks, the officials said.

This so-called "pattern analysis" on calls within the United States would, in many circumstances, require a court warrant if the government wanted to trace who calls whom.

The use of similar data-mining operations by the Bush administration in other contexts has raised strong objections, most notably in connection with the Total Information Awareness system, developed by the Pentagon for tracking terror suspects, and the Department of Homeland Security's Capps program for screening airline passengers. Both programs were ultimately scrapped after public outcries over possible threats to privacy and civil liberties.

But the Bush administration regards the N.S.A.'s ability to trace and analyze large volumes of data as critical to its expanded mission to detect terrorist plots before they can be carried out, officials familiar with the program say. Administration officials maintain that the system set up by Congress in 1978 under the Foreign Intelligence Surveillance Act does not give them the speed and flexibility to respond fully to terrorist threats at home.

A former technology manager at a major telecommunications company said that since the Sept. 11 attacks, the leading companies in the industry have been storing information on calling patterns and giving it to the federal government to aid in tracking possible terrorists.

"All that data is mined with the cooperation of the government and shared with them, and since 9/11, there's been much more active involvement in that area," said the former manager, a telecommunications expert who did not want his name or that of his former company used because of concern about revealing trade secrets.

Such information often proves just as valuable to the government as eavesdropping on the calls themselves, the former manager said.

"If they get content, that's useful to them too, but the real plum is going to be the transaction data and the traffic analysis," he said. "Massive amounts of traffic analysis information -- who is calling whom, who is in Osama Bin Laden's circle of family and friends -- is used to identify lines of communication that are then given closer scrutiny."

Several officials said that after President Bush's order authorizing the N.S.A. program, senior government officials arranged with officials of some of the nation's largest telecommunications companies to gain access to switches that act as gateways at the borders between the United States' communications networks and international networks. The identities of the corporations involved could not be determined.

The switches are some of the main arteries for moving voice and some Internet traffic into and out of the United States, and, with the globalization of the telecommunications industry in recent years, many international-to-international calls are also routed through such American switches.

One outside expert on communications privacy who previously worked at the N.S.A. said that to exploit its technological capabilities, the American government had in the last few years been quietly encouraging the telecommunications industry to increase the amount of international traffic that is routed through American-based switches.

The growth of that transit traffic had become a major issue for the intelligence community, officials say, because it had not been fully addressed by 1970's-era laws and regulations governing the N.S.A. Now that foreign calls were being routed through switches on American soil, some judges and law enforcement officials regarded eavesdropping on those calls as a possible violation of those decades-old restrictions, including the Foreign Intelligence Surveillance Act, which requires court-approved warrants for domestic surveillance.

Historically, the American intelligence community has had close relationships with many communications and computer firms and related technical industries. But the N.S.A.'s backdoor access to major telecommunications switches on American soil with the cooperation of major corporations represents a significant expansion of the agency's operational capability, according to current and former government officials.

Phil Karn, a computer engineer and technology expert at a major West Coast telecommunications company, said access to such switches would be significant. "If the government is gaining access to the switches like this, what you're really talking about is the capability of an enormous vacuum operation to sweep up data," he said.

Copyright 2006 The New York Times Company | Privacy Policy | Home | Search | Corrections | Help | Back to Top

Exhibit C

washingtonpost.com

Surveillance Net Yields Few Suspects

NSA's Hunt for Terrorists Scrutinizes Thousands of Americans, but Most Are Later Cleared

By Barton Gellman, Dafna Linzer and Carol D. Leonnig
Washington Post Staff Writers
Sunday, February 5, 2006; A01

Advertisement

CONVERSATIONS
WITH MICHAEL EISNER
FREEWHEELING INTERVIEWS WITH LEADING INNOVATORS FROM BUSINESS AND ENTERTAINMENT

Intelligence officers who eavesdropped on thousands of Americans in overseas calls under authority from President Bush have dismissed nearly all of them as potential suspects after hearing nothing pertinent to a terrorist threat, according to accounts from current and former government officials and private-sector sources with knowledge of the technologies in use.

Bush has recently described the warrantless operation as "terrorist surveillance" and summed it up by declaring that "if you're talking to a member of al Qaeda, we want to know why." But officials conversant with the program said a far more common question for eavesdroppers is whether, not why, a terrorist plotter is on either end of the call. The answer, they said, is usually no.

Fewer than 10 U.S. citizens or residents a year, according to an authoritative account, have aroused enough suspicion during warrantless eavesdropping to justify interception of their domestic calls, as well. That step still requires a warrant from a federal judge, for which the government must supply evidence of probable cause.

The Bush administration refuses to say -- in public or in closed session of Congress -- how many Americans in the past four years have had their conversations recorded or their e-mails read by intelligence analysts without court authority. Two knowledgeable sources placed that number in the thousands; one of them, more specific, said about 5,000.

The program has touched many more Americans than that. Surveillance takes place in several stages, officials said, the earliest by machine. Computer-controlled systems collect and sift basic information about hundreds of thousands of faxes, e-mails and telephone calls into and out of the United States before selecting the ones for scrutiny by human eyes and ears.

Successive stages of filtering grow more intrusive as artificial intelligence systems rank voice and data traffic in order of likeliest interest to human analysts. But intelligence officers, who test the computer judgments by listening initially to brief fragments of conversation, "wash out" most of the leads within days or weeks.

The scale of warrantless surveillance, and the high proportion of bystanders swept in, sheds new light on Bush's circumvention of the courts. National security lawyers, in and out of government, said the washout rate raised fresh doubts about the program's lawfulness under the Fourth Amendment, because a search cannot be judged "reasonable" if it is based on evidence that experience shows to be unreliable. Other officials said the disclosures might shift the terms of public debate, altering perceptions about the balance between privacy lost and security gained.

Air Force Gen. Michael V. Hayden, the nation's second-ranking intelligence officer, acknowledged in a news briefing last month that eavesdroppers "have to go down some blind alleys to find the tips that pay off." Other officials, nearly all of whom spoke on the condition of anonymity because they are not permitted to discuss the program, said the prevalence of false leads is especially pronounced when U.S. citizens or residents are surveilled. No intelligence agency, they said, believes that "terrorist . . . operatives inside our country," as Bush described the surveillance targets, number anywhere near the thousands who have been subject to eavesdropping.

The Bush administration declined to address the washout rate or answer any other question for this article about the policies and operations of its warrantless eavesdropping.

Vice President Cheney has made the administration's strongest claim about the program's intelligence value, telling CNN in December that eavesdropping without warrants "has saved thousands of lives." Asked about that Thursday, Hayden told senators he "cannot personally estimate" such a figure but that the program supplied information "that would not otherwise have been available." FBI Director Robert S. Mueller III said at the same hearing that the information helped identify "individuals who were providing material support to terrorists."

Supporters speaking unofficially said the program is designed to warn of unexpected threats, and they argued that success cannot be measured by the number of suspects it confirms. Even unwitting Americans, they said, can take part in communications -- arranging a car rental, for example, without knowing its purpose -- that supply "indications and warnings" of an attack. Contributors to the technology said it is a triumph for artificial intelligence if a fraction of 1 percent of the computer-flagged conversations guide human analysts to meaningful leads.

Those arguments point to a conflict between the program's operational aims and the legal and political limits described by the president and his advisers. For purposes of threat detection, officials said, the analysis of a telephone call is indifferent to whether an American is on the line. Since Sept. 11, 2001, a former CIA official said, "there is a lot of discussion" among analysts "that we shouldn't be dividing Americans and foreigners, but terrorists and non-terrorists." But under the Constitution, and in the Bush administration's portrait of its warrantless eavesdropping, the distinction is fundamental.

Valuable information remains valuable even if it comes from one in a thousand intercepts. But government officials and lawyers said the ratio of success to failure matters greatly when eavesdropping subjects are Americans or U.S. visitors with constitutional protection. The minimum legal definition of probable cause, said a government official who has studied the program closely, is that evidence used to support eavesdropping ought to turn out to be "right for one out of every two guys at least." Those who devised the surveillance plan, the official said, "knew they could never meet that standard -- that's why they didn't go through" the court that supervises the Foreign Intelligence Surveillance Act, or FISA.

Michael J. Woods, who was chief of the FBI's national security law unit until 2002, said in an e-mail interview that even using the lesser standard of a "reasonable basis" requires evidence "that would lead a prudent, appropriately experienced person" to believe the American is a terrorist agent. If a factor returned "a large number of false positives, I would have to conclude that the factor is not a sufficiently reliable indicator and thus would carry less (or no) weight."

Bush has said his program covers only overseas calls to or from the United States and stated categorically that "we will not listen inside this country" without a warrant. Hayden said the government goes to the intelligence court when an eavesdropping subject becomes important enough to "drill down," as he put it, "to the degree that we need all communications."

Yet a special channel set up for just that purpose four years ago has gone largely unused, according to an authoritative account. Since early 2002, when the presiding judge of the federal intelligence court first learned of Bush's program, he agreed to a system in which prosecutors may apply for a domestic warrant after warrantless eavesdropping on the same person's overseas communications. The annual number of such applications, a source said, has been in the single digits.

Many features of the surveillance program remain unknown, including what becomes of the non-threatening U.S. e-mails and conversations that the NSA intercepts. Participants, according to a national security lawyer who represents one of them privately, are growing "uncomfortable with the mountain of data they have now begun to accumulate." Spokesmen for the Bush administration declined to say whether any are discarded.

New Imperatives

Recent interviews have described the program's origins after Sept. 11 in what Hayden has called a three-way collision of "operational, technical and legal imperatives."

Intelligence agencies had an urgent mission to find hidden plotters before they could strike again.

About the same time, advances in technology -- involving acoustic engineering, statistical theory and efficient use of computing power to apply them -- offered new hope of plucking valuable messages from the vast flow of global voice and data traffic. And rapidly changing commercial trends, which had worked against the NSA in the 1990s as traffic shifted from satellites to fiber-optic cable, now presented the eavesdroppers with a gift. Market forces were steering as much as a third of global communications traffic on routes that passed through the United States.

The Bush administration had incentive and capabilities for a new kind of espionage, but 23 years of law and White House policy stood in the way.

FISA, passed in 1978, was ambiguous about some of the president's plans, according to current and retired government national security lawyers. But other features of the eavesdropping program fell outside its boundaries.

One thing the NSA wanted was access to the growing fraction of global telecommunications that passed through junctions on U.S. territory. According to former senator Bob Graham (D-Fla.), who chaired the Intelligence Committee at the time, briefers told him in Cheney's office in October 2002 that Bush had authorized the agency to tap into those junctions. That decision, Graham said in an interview first reported in The Washington Post on Dec. 18, allowed the NSA to intercept "conversations that . . . went through a transit facility inside the United States."

According to surveys by TeleGeography Inc., nearly all voice and data traffic to and from the United States now travels by fiber-optic cable. About one-third of that volume is in transit from one foreign country to another, traversing U.S. networks along its route. The traffic passes through cable landing stations, where undersea communications lines meet the East and West coasts; warehouse-size gateways where competing international carriers join their networks; and major Internet hubs known as metropolitan area ethernets.

Until Bush secretly changed the rules, the government could not tap into access points on U.S. soil without a warrant to collect the "contents" of any communication "to or from a person in the United States." But the FISA law was silent on calls and e-mails that began and ended abroad.

Even for U.S. communications, the law was less than clear about whether the NSA could harvest information about that communication that was not part of its "contents."

"We debated a lot of issues involving the 'metadata,' " one government lawyer said. Valuable for analyzing calling patterns, the metadata for telephone calls identify their origin, destination, duration and time. E-mail headers carry much the same information, along with the numeric address of each network switch through which a message has passed.

Intelligence lawyers said FISA plainly requires a warrant if the government wants real-time access to that information for any one person at a time. But the FISA court, as some lawyers saw it, had no explicit jurisdiction over wholesale collection of records that do not include the content of communications. One high-ranking intelligence official who argued for a more cautious approach said he found himself pushed aside. Awkward silences began to intrude on meetings that discussed the evolving rules.

"I became aware at some point of things I was not being told about," the intelligence official said.

'Subtly Softer Trigger'

Hayden has described a "subtly softer trigger" for eavesdropping, based on a powerful "line of logic," but no Bush administration official has acknowledged explicitly that automated filters play a role in selecting American targets. But Sen. Arlen Specter (R-Pa.), who chairs the Judiciary Committee, referred in a recent letter to "mechanical surveillance" that is taking place before U.S. citizens and residents are "subject to human surveillance."

Machine selection would be simple if the typical U.S. eavesdropping subject took part in direct calls to or from the "phone numbers of known al Qaeda" terrorists, the only criterion Bush has mentioned.

That is unusual. The NSA more commonly looks for less-obvious clues in the "terabytes of speech, text, and image data" that its global operations collect each day, according to an unclassified report by the National Science Foundation soliciting research on behalf of U.S. intelligence.

NSA Inspector General Joel F. Brenner said in 2004 that the agency's intelligence officers have no choice but to rely on "electronic filtering, sorting and dissemination systems of amazing sophistication but that are imperfect."

One method in use, the NSF report said, is "link analysis." It takes an established starting point -- such as a terrorist just captured or killed -- and looks for associated people, places, things and events. Those links can be far more tenuous than they initially appear.

In an unclassified report for the Pentagon's since-abandoned Total Information Awareness program, consultant Mary DeRosa showed how "degrees of separation" among the Sept. 11 conspirators concealed the significance of clues that linked them.

Khalid Almihdhar, one of the hijackers, was on a government watch list for terrorists and thus a known suspect. Mohamed Atta, another hijacker, was linked to Almihdhar by one degree of separation because he used the same contact address when booking his flight. Wail M. Alshehri, another hijacker, was linked by two degrees of separation because he shared a telephone number with Atta. Satam M.A. Al Suqami, still another hijacker, shared a post office box with Alshehri and, therefore, had three degrees of separation from the original suspect.

'Look for Patterns'

Those links were not obvious before the identity of the hijackers became known. A major problem for analysts is that a given suspect may have hundreds of links to others with one degree of separation, including high school classmates and former neighbors in a high-rise building who never knew his name. Most people are linked to thousands or tens of thousands of people by two degrees of separation, and hundreds of thousands or millions by three degrees.

Published government reports say the NSA and other data miners use mathematical techniques to form hypotheses about which of the countless theoretical ties are likeliest to represent a real-world relationship.

A more fundamental problem, according to a high-ranking former official with firsthand knowledge, is that "the number of identifiable terrorist entities is decreasing." There are fewer starting points, he said, for link analysis.

"At that point, your only recourse is to look for patterns," the official said.

Pattern analysis, also described in the NSF and DeRosa reports, does not depend on ties to a known suspect. It begins with places terrorists go, such as the Pakistani province of Waziristan, and things they do, such as using disposable cell phones and changing them frequently, which U.S. officials have publicly cited as a challenge for counterterrorism.

"These people don't want to be on the phone too long," said Russell Tice, a former NSA analyst, offering another example.

Analysts build a model of hypothetical terrorist behavior, and computers look for people who fit the model. Among the drawbacks of this method is that nearly all its selection criteria are innocent on their own. There is little precedent, lawyers said, for using such a model as probable cause to get a court-issued warrant for electronic surveillance.

Jeff Jonas, now chief scientist at IBM Entity Analytics, invented a data-mining technology used widely in the private sector and by the government. He sympathizes, he said, with an analyst facing an unknown threat who gathers enormous volumes of data "and says, 'There must be a secret in there.' "

But pattern matching, he argued, will not find it. Techniques that "look at people's behavior to predict terrorist intent," he said, "are so far from reaching the level of accuracy that's necessary that I see them as nothing but civil liberty infringement engines."

'A Lot Better Than Chance'

Even with 38,000 employees, the NSA is incapable of translating, transcribing and analyzing more than a fraction of the conversations it intercepts. For years, including in public testimony by Hayden, the agency has acknowledged use of automated equipment to analyze the contents and guide analysts to the most important ones.

According to one knowledgeable source, the warrantless program also uses those methods. That is significant to the public debate because this kind of filtering intrudes into content, and machines "listen" to more Americans than humans do. NSA rules since the late 1970s, when machine filtering was far less capable, have said "acquisition" of content does not take place until a conversation is intercepted and processed "into an intelligible form intended for human inspection."

The agency's filters are capable of comparing spoken language to a "dictionary" of key words, but Roger W. Cressey, a senior White House counterterrorism official until late 2002, said terrorists and other surveillance subjects make frequent changes in their code words. He said, " 'Wedding' was martyrdom day and the 'bride' and 'groom' were the martyrs." But al Qaeda has stopped using those codes.

An alternative approach, in which a knowledgeable source said the NSA's work parallels academic and commercial counterparts, relies on "decomposing an audio signal" to find qualities useful to pattern analysis. Among the fields involved are acoustic engineering, behavioral psychology and computational linguistics.

A published report for the Defense Advanced Research Projects Agency said machines can easily determine the sex, approximate age and social class of a speaker. They are also learning to look for clues to deceptive intent in the words and "paralinguistic" features of a conversation, such as pitch, tone, cadence and latency.

This kind of analysis can predict with results "a hell of a lot better than chance" the likelihood that the speakers are trying to conceal their true meaning, according to James W. Pennebaker, who chairs the psychology department at the University of Texas at Austin.

"Frankly, we'll probably be wrong 99 percent of the time," he said, "but 1 percent is far better than 1 in 100 million times if you were just guessing at random. And this is where the culture has to make some decisions."

Researcher Julie Tate and staff writer R. Jeffrey Smith contributed to this report.

© 2006 The Washington Post Company

Ads by Google

Is Bush Doing A Good Job?
Take Our Poll And You Can Get A Free Laptop Computer
www.will-bush-win.com

George Bush Bill Clinton
Who Is A Better President? Vote Now To See Who's Winning!
www.popularq.com

Why did Bush avoid FISA?
Tell Congress we need the facts about Bush's domestic surveillance
UniteOurStates.com

Exhibit D

RAND Journal of Economics
Vol. 34, No. 2, Summer 2003
pp. 370–390

Internet interconnection and the off-net-cost pricing principle

Jean-Jacques Laffont*

Scott Marcus**

Patrick Rey***

and

Jean Tirole****

We develop a framework for Internet backbone competition. In the absence of direct payments between websites and consumers, the access charge allocates communication costs between websites and consumers and affects the volume of traffic. We analyze the impact of the access charge on competitive strategies in an unregulated retail environment. In a remarkably broad range of environments, operators set prices for their customers as if their customers' traffic were entirely off-net. We then compare the socially optimal access charge with the privately desirable one. Finally, when websites charge micropayments, or sell goods and services, the impact of the access charge on welfare is reduced; in particular, the access charge is neutral in a range of circumstances.

1. Introduction

■ Long an emanation of voluntarist public policies, the Internet has moved in recent years to a market paradigm. While still partly run on the basis of legacy agreements, the Internet industry is actively searching for a business model that will increase Internet usage and facilitate the evolution to enhanced offerings based on differentiated classes of services. A key feature of the Internet is that each computer connected to it can communicate with every other connected computer. In a deregulated environment, this universal connectivity can be achieved only if competing connectivity providers cooperatively reach agreements governing the price and quality of their interconnection.

* University of Toulouse (IDEI and GREMAQ); laffont@cict.fr.

** U.S. Federal Communications Commission; scott@scottmarcus.com.

*** University of Toulouse (IDEI and GREMAQ; prey@cict.fr.

**** University of Toulouse (IDEI and GREMAQ), CERAS (Paris), and MIT; tirole@cict.fr.

We are grateful to Mike Riordan, Aaron Schiff, and Julian Wright, and to the participants of the IDEI conference on the Economics of the Software and Internet Industries (January 18–20, 2001) for helpful reactions on an earlier draft. Concerning Scott Marcus, the opinions expressed do not necessarily reflect the views of the FCC or any of its commissioners. We also thank the Editor, Jennifer Reinganum, and two referees for their comments.

Copyright © 2003, RAND.

The interconnection charges, also called "access charges," "settlements," or "termination charges," could be vital for enabling efficient use of the Internet. Incentives must be provided for a widespread usage of bandwidth by dial-up, broadband, and dedicated access consumers, and for the posting of content by the websites. Quality-of-service (QoS) agreements between operators can reduce delays and packet losses for marked traffic and thereby enable the development of new and advanced Internet services such as IP telephony and videoconferencing. Competition for end users is a necessary condition for an efficient functioning of the industry, but it will fall short of accomplishing even its most modest goals in the absence of proper interconnection agreements.

The purpose of this article is to develop a framework for modelling the competition among interconnected Internet "backbone operators" or "networks." In this framework, the "end users" or "customers" are heterogeneous in several respects. First, their patterns of traffic imbalance differ. Consumers receive much more traffic than they send, primarily due to the downloads they request; websites, in contrast, originate much of their traffic, even though they do not request it. Second, different end users generate different value to other end users and thus to the Internet. Third, end users may differ in the cost that their traffic imposes on the operators.

The backbone operators vie for the various types of traffic. In particular, each competes on the two sides of the market (consumers and websites). The competitive analysis offers two sets of insights:

Competitive strategies. On the positive side, we analyze pricing strategies in this interconnected environment. The first key insight of the article is that in a wide range of situations, backbones set their price on each business segment *as if* they had no other customer. That is, they set charges to consumers and websites as if their connections were entirely off-net. We call this the "off-net-cost pricing principle." We first demonstrate this principle in the simplest perfectly competitive environment with a reciprocal access charge. This simple principle turns out to be remarkably robust to generalizations of the model: mixed traffic patterns, variable demand, QoS agreements, backbone differentiation, installed bases, multihoming, and customer cost heterogeneity.

Impact of the access charge on welfare and profit. The access charge affects the backbones' marginal cost of incoming and outgoing off-net traffic. It therefore determines how backbones distribute communication costs between websites and consumers. *Ceteris paribus,* a higher access charge penalizes end users, such as websites, with an outgoing-traffic bias, and it benefits end users, such as consumers, with the opposite bias. Network externalities considerations, though, complicate end users' preferences over access charges, as they want the other side of the market to expand.

We first consider the case where there is no direct payment between websites and consumers. This case is most relevant when there are no micropayments and no other financial transaction resulting from consumers' visits to the websites. In that case, the access charge should promote economic efficiency by alleviating the burden on those end users (i) whose demand is highly elastic and (ii) who create value for other end users. More generally, we shall argue that the access charge cannot by itself induce all the price differentiation that would be required for an efficient allocation in the Internet. Furthermore, if backbones have market power, they do not necessarily choose the socially optimal access charge.

Also, individual end users' elasticities will be affected by a more widespread use of micropayments between end users, which partly reallocate costs endogenously. Indeed, we consider more briefly the case where consumers pay a price to the websites for their visits (this price can be a micropayment charged by the website, or part of a transaction resulting from their visit). This financial transaction provides an additional channel for allocating the cost of the communication, which lowers the allocative impact of the access charge.

On the positive side, we analyze the access charge's impact on profits. There may be no such impact, for example when an increase in the access charge is competed away by the backbones' offering very low prices to consumers. If backbones have market power, however, profits are affected by the access charge, and backbones will tend to subsidize the more profitable segment.

© RAND 2003.

The article is organized as follows. Section 2 constructs a model of perfect (Bertrand) backbone competition for consumers and websites, assuming that both sides of the market are supplied, i.e., demands are locally inelastic. Section 3 demonstrates the robustness of the off-net-cost pricing principle. Section 4 analyzes the socially optimal access charge. Section 5 discusses some limits of the off-net-cost pricing principle. Section 6 introduces micropayments between customers and websites. Section 7 concludes.

Our article is related to the literature on two-way access in telecommunications, e.g., Armstrong (1998) and Laffont, Rey, and Tirole (1998a 1998b).[1] This literature assumes that while consumers both send and receive traffic, receivers get no surplus from and are not charged for receiving calls. When instead receivers derive some utility from receiving calls, an externality must be internalized for efficiency. The fact that users are not charged for receiving traffic has several implications. First, operators' marginal charge for outgoing traffic is equal to the on-net cost augmented by the *average* termination markup rather than to the off-net cost. Second, it creates some instability in competition if the networks are close substitutes and the termination charge is not in the vicinity of the termination cost; in contrast, Section 2 establishes that no such instability occurs when consumers are charged for receiving calls.

The articles most related to ours are Jeon, Laffont, and Tirole (2001) and Hermalin and Katz (2001). Jeon, Laffont, and Tirole analyze the off-net cost pricing principle in a telecommunications environment where the volume of traffic between each sender and receiver is endogenously determined by the party with the lower marginal willingness to communicate. In particular, this formulation allows one to tell apart monthly (subscription) fees and usage fees (for receiving and sending traffic). That article also considers the case of regulated reception charges, and it stresses furthermore that network-based price discrimination is conducive to connectivity breakdowns. In contrast, in most of this article we suppose that there is a fixed volume of transactions for each consumer-website match. This makes nonlinear tariffs irrelevant (no distinction between fixed and usage fees); we use this simpler formulation to study several additional aspects, such as the impact of multihoming, market power, asymmetric access charges, and micropayments between consumers and websites. Hermalin and Katz also focus on fixed transactions but allow for stochastic (and possibly correlated) gains from communication. They show that double marginalization increases when networks specialize in offering services to senders or receivers and also study asymmetric Bertrand competition, where some operators are more efficient than others.

2. A simple benchmark

■ Although our theory allows for general traffic imbalances, it is useful for expository purposes to distinguish two types of customers: *websites* and *consumers*. Consumers exchange traffic (e.g., emails), browse web pages, download files, and so forth; websites post pages and files, which can be browsed and downloaded by consumers. There is little traffic between websites, and furthermore, the traffic between consumers (such as email exchanges) or from consumers to websites (the requests for pages or file downloads) is much smaller than the traffic from websites to consumers (the actual downloading of web pages and files). To capture this traffic pattern in its simplest form, we neglect the traffic between consumers or between websites, as well as the traffic from consumers to websites, and focus instead on the traffic from websites to consumers.

Most of the article uses the following assumptions:

Balanced calling pattern. We assume that consumers' interest in a website is unrelated to the website's network choice: a consumer is as likely to request a page from a given website belonging to her network and another given website belonging to a rival network.[2] In the absence of

[1] See also Carter and Wright (1999a, 1999b), Cherdron (2000), Dessein (forthcoming), Gans and King (2001), and Hahn (2000).

[2] This assumption ought to be refined in specific instances. For example, regional or international specialization of backbones together with other factors, such as language affinity, may induce some violations of this hypothesis (Chinese consumers may be more likely to browse U.S. websites than U.S. customers to browse Chinese websites).

© RAND 2003.

origination-based price discrimination (that is, if a consumer pays the same price for receiving traffic, regardless of the identity of the originating website's backbone), the percentage of traffic originating on network i and completed on network j is therefore proportional both to the fraction of websites on network j and to the fraction of consumers subscribing to network i.

Reciprocal access pricing. We assume that there is no asymmetry in the interconnection charge: A network pays as much for having its traffic terminated on a rival network ("off-net traffic") as it receives for terminating traffic originating on a rival network. This assumption will be relaxed in Section 5, but it is worth noting that there have been calls for regulators to impose reciprocal access charges.[3] (At the moment, most interconnection agreements between the top-level backbones take the form of "bill and keep" peering agreements, with zero (and thus reciprocal) termination charges; however, this situation is likely to evolve in the future—some backbones have already introduced positive termination charges in their agreements with certain other backbones.)

Let us now be more specific about the model:

Cost structure. Two full-coverage[4] "networks," or "backbones" or "operators," have the same cost structure. For notational simplicity, we ignore request traffic, so that the only costs are those incurred to bring traffic from websites to consumers. We also do not include any fixed network cost. It is straightforward to add both types of costs.[5]

We let c denote the total marginal cost of traffic. When traffic is handed over from one backbone to the other, we let c_o and c_t denote the originating and terminating backbones' marginal costs associated with this traffic ($c_o + c_t = c$).

Although the exact expressions of c_o and c_t are irrelevant for the theory, it is useful for concreteness to discuss the nature of these costs in the current Internet environment. For example, suppose that backbones incur a marginal cost c' per unit of traffic at the originating and terminating ends and a marginal cost c'' in between, which may stand for the routing costs and the marginal cost of trunk lines used for transportation. The total marginal cost of traffic is thus

$$c \equiv 2c' + c''.$$

In practice, top-level backbone operators have multiple interconnection points and have an incentive to pass on off-net traffic as soon as possible. A consequence of this "hot-potato" pattern[6] is that most of the transportation cost c'' is borne by the receiving backbone.[7] For off-net traffic, the sending network thus incurs the marginal cost of origination, c', while the receiving network incurs both the transportation cost c'' and the marginal cost of termination, c'. The total marginal cost of traffic is thus shared by the sending and receiving networks according to

$$c_o \equiv c' \quad \text{and} \quad c_t \equiv c' + c''.$$

Demand structure. We first assume that the networks are perfect substitutes: that consumers and websites have inelastic demand for and supply of web pages. To be sure, consumers and websites

---

[3] See Marcus (1999) and Gao (2000) for overviews of the Internet's hierarchical organization.

[4] "Full coverage" means that the backbones have a global geographical presence and thus are able to serve all customers.

[5] The next section considers mixed traffic patterns. For simplicity, we also ignore the impact on the cost structure of caching, replication, and other content-delivery network schemes.

[6] For a description of hot-potato routing, see Marcus (1999).

[7] Our analysis would, however, apply to any other method of sharing the cost of off-net traffic. We assume here that the access charge is, as is currently the case, independent of the "distance" between the point at which the traffic is handed over and the location of the receiver. Our analysis would still apply if there were differentiated access charges, as long as differences in access charges reflected differences in termination costs. The white paper NRIC (2002) provides a detailed overview of current interconnection agreements and the issues they raise.

© RAND 2003.

are more likely to use the web if they are charged lower prices; we will thus relax these assumptions later on.

There is a continuum of consumers, of mass 1, and a continuum of websites, of mass 1 as well. Each consumer generates one unit of traffic from each website connected to either backbone. Each unit of traffic from a website to a consumer yields a value v to the consumer and a value $\tilde{v}$ to the website. We will assume that the market is viable, that is,

$$v + \tilde{v} > c.$$

Until Section 6, we assume away "micropayments" between consumers and websites and posit that websites do not charge differentiated prices to consumers depending on whether their connection is on- or off-net. Furthermore, backbones are perfect substitutes on both sides of the market, and so each side chooses the lowest price that it is offered.

We will initially assume that prices are low enough that all consumers or websites connect to a backbone. The volume of traffic associated with each customer is then fixed, and there is thus no point distinguishing between subscription and usage prices or linear and nonlinear prices: consumers' subscription decisions are based on the prices p_1 and p_2 charged by the two backbones for receiving traffic, while websites' subscription decisions are based on the prices $\tilde{p}_1$ and $\tilde{p}_2$ charged for sending traffic.[8] Note that the backbones need not be able to tell consumers and websites apart directly. It suffices that inflows and outflows be priced differently.

Denoting by α_i backbone i's market share for consumers and by $\tilde{\alpha}_i$ its market share for websites, and assuming that the two operators charge each other the same interconnection charge a for terminating traffic, backbone i's profit is given by (for $i \neq j = 1, 2$):

$$\pi_i = \alpha_i \tilde{\alpha}_i \left(p_i + \tilde{p}_i - c\right) + \alpha_i \tilde{\alpha}_j \left(p_i - (c_t - a)\right) + \alpha_j \tilde{\alpha}_i \left(\tilde{p}_i - (c_o + a)\right)$$

or

$$\pi_i = \alpha_i \tilde{\alpha} \left[p_i - (c_t - a)\right] + \tilde{\alpha}_i \alpha \left[\tilde{p}_i - (c_o + a)\right], \tag{1}$$

where $\alpha = \alpha_1 + \alpha_2$ and $\tilde{\alpha} = \tilde{\alpha}_1 + \tilde{\alpha}_2$ denote, respectively, the total numbers of connected consumers and of connected websites. If all potential customers are connected as we assume in this section (that is, $\alpha = \tilde{\alpha} = 1$), this expression reduces to

$$\pi_i = \alpha_i \left[p_i - (c_t - a)\right] + \tilde{\alpha}_i \left[\tilde{p}_i - (c_o + a)\right]. \tag{2}$$

That is, as long as prices do not exceed customers' reservation values, the profit of each backbone can be decomposed into two independent components: one for the consumer business, and another one for the website business. The perfect-substitutability assumption ensures furthermore that, in each line of business, all customers go to the cheapest operator whenever their prices differ.

The timing is as follows: (1) the access charge a is determined (through a bilateral agreement or by regulation), (2) the backbones set their prices, and (3) end users select their backbones. As is usual, we solve for a subgame-perfect equilibrium of the game.

Proposition 1 (off-net-cost pricing principle). Assume $v \geq c_t - a$ and $\tilde{v} \geq c_o + a$; then, there exists a unique price equilibrium.[9] This equilibrium is symmetric and satisfies

$$p_1 = p_2 = p^* = c_t - a,$$
$$\tilde{p}_1 = \tilde{p}_2 = \tilde{p}^* = c_o + a,$$

[8] The consumer prices p_1 and p_2 can be indifferently interpreted as volume-based prices for receiving traffic, or as subscription prices—if the total number of websites were not normalized to 1, these would be subscription prices per website reached. Similarly, websites' prices $\tilde{p}_1$ and $\tilde{p}_2$ can be interpreted as subscription prices (per consumer reached).

[9] Market shares are undetermined.

© RAND 2003.

$$\pi_1 = \pi_2 = \pi^* = 0.$$

Proof. The standard Bertrand argument applies to each business segment. The only caveat is that the number of connected customers in one segment affects the market demand in the other segment; however, as long as prices remain below reservation values, all customers are connected (to one or the other network) and, in each segment, the market demand is thus independent of the actual price levels. *Q.E.D.*

For each customer, the price is competitively set equal to the opportunity cost of servicing this customer, rather than letting the customer subscribe to the other network. Suppose for example that backbone 1 "steals" a consumer away from backbone 2. Then, the traffic from backbone 2's websites to that consumer, which was previously internal to backbone 2, now costs backbone 1 an amount c_t to terminate but generates a marginal termination revenue a; the opportunity cost of that traffic is thus $c_t - a$. And the traffic from backbone 1's websites, which initially costs c_o for origination and a for termination on backbone 2, is now internal to backbone 1 and thus costs $c = c_o + c_t$; therefore, for that traffic too, the opportunity cost of stealing the consumer away from its rival is $c - (c_o + a) = c_t - a$. A similar reasoning shows that stealing a website away from the rival backbone generates, for each connected consumer, a net cost $c_o + a$: attracting a website increases originating traffic, which costs c_o, and also means sending more traffic from its own websites to the other backbone's end users, as well as receiving less traffic from the other backbone (since the traffic originated by the stolen backbone is now on-net); in both cases, a termination revenue a is lost.

In this very simple benchmark case of perfectly substitutable networks and inelastic demand, Bertrand-like competition ensures that profits are set at their competitive level ($\pi^* = 0$); whatever the access charge a, the combined per-unit charge to consumers and websites covers the cost of the traffic:[10]

$$\tilde{p}^* + p^* = (c_o + a) + (c_t - a) = c_o + c_t = c.$$

The access charge a thus merely determines how the cost of the traffic is shared between senders (websites) and receivers (consumers)—with a higher access charge leading to a larger burden being placed on the websites. In particular, the access charge has no impact on network profits and on social welfare, defined as the sum of customers' surpluses, which is equal to its first-best level:

$$\begin{aligned} W &= \Sigma_i \alpha_i \tilde{\alpha} (v - p_i) + \Sigma_i \tilde{\alpha}_i \alpha (\tilde{v} - \tilde{p}_i) + \Sigma_i \pi_i \\ &= W^{FB} \equiv v + \tilde{v} - c. \end{aligned}$$

Finally, let us compare Proposition 1 with the results in Laffont, Rey, and Tirole (1998a) and Armstrong (1998) for interconnection of telephone networks. A key difference with this telecommunications literature is that in the latter there is a missing price: receivers do not pay for receiving calls; that is, in the notation of this article, $p = 0$. The missing price has two important implications:

Pricing. The operators' optimal usage price reflects their *perceived* marginal cost. But when operators do not charge their customers (here, consumers) for the traffic they receive, operator i's perceived marginal cost of outgoing (here, website) traffic is given by

$$c + \alpha_j (a - c_t). \tag{3}$$

That is, the unit cost of traffic is the on-net cost c, augmented by the expected off-net "markup"

[10] This holds as long as customers' prices remain lower than customers' reservation values, that is, as long as $c_o + a \leq \tilde{v}$ and $c_t - a \leq v$. If for example $c_o + a > \tilde{v}$, the maximal price that can be charged to websites, $\tilde{p} = \tilde{v}$, does not cover the opportunity cost they generate, $c_o + a$. Thus, no backbone wants to host a website and there is then no traffic at all for such an access charge.

© RAND 2003.

(or discount) $(a - c_t)$ on the fraction α_j of website traffic that terminates off-net. Comparing the two perceived marginal costs of outgoing traffic with and without receiver charge, for given access charge and market shares, the price for sending traffic is higher (lower) than in the presence of reception charges if and only if there is a termination discount (markup).[11]

Note that if the "missing payment" $\alpha_i p_i$ were subtracted from the right-hand side[12] of (3) and p_i were equal to the off-net cost[13] $(c_t - a)$, then (3) would be equal to the off-net cost $(c_o + a)$. In sum, the missing payment affects the backbones' perceived costs, and it reallocates costs between origination and reception.

Stability in competition. When networks are close substitutes, and receivers are not charged, there exists no equilibrium unless the access charge is near the termination cost. The intuition is easily grasped from (3). If there is a substantial termination tax or subsidy, perceived marginal costs (and thus prices) are far from actual costs, thereby introducing a source of inefficiency. But if networks are sufficiently close substitutes, either operator could corner the market with a small reduction in its price, in which case it faces the true costs and can offer a better deal. This issue does not arise when end users pay (or are paid) for receiving traffic. In that case, the sum of the perceived costs for origination and termination always equals the actual cost of communication: $(c_o + a) + (c_t - a) = c$, irrespective of the access charge.

3. Robustness of the off-net-cost pricing principle

■ The off-net-cost pricing principle is robust to various extensions of the perfectly competitive model.

(i) *Arbitrary number of backbones.* The principle extends trivially to n backbones ($n \geq 2$): it suffices to replace "α_j" in equation (1) by "$\sum_{j \neq i} \alpha_j$".

(ii) *Mixed traffic patterns.* We have caricatured reality by assuming that websites have only outgoing traffic, and consumers only incoming traffic. All Internet users in fact have a mixed, although often very biased, pattern. It is easily verified that under perfect competition, backbones ask their customers (consumers or websites) to pay

$$T_i(x, y) = (c_t - a)x + (c_o + a)y,$$

where x and y are the customer's incoming and outgoing traffic volumes.

(iii) *Multihoming.* Suppose now that each website may choose to locate in both backbones. Websites do not gain or lose from multihoming as long as the backbones charge the competitive tariff $\tilde{p}^* = c_o + a$.[14]

(iv) *Quality of service (QoS).* Proposition 1 extends to multiple qualities of service, as long as costs and access charge refer to the quality of service in question.

(v) *Customer cost heterogeneity.* Our assumption that all customers impose the same cost on the backbone for incoming or outgoing traffic is more restrictive than needed. Suppose that there are K types of customers, $k = 1, \ldots, K$. A customer of type k, whether a consumer or a website, imposes cost c_o^k at origination and c_t^k at termination.[15] The off-net-cost pricing principle still holds as long as backbones can price discriminate.

[11] Indeed, $c + \alpha_j(a - c_t) > c_o + a$ is equivalent to $(1 - \alpha_j)(a - c_t) < 0$.

[12] To reflect the fact that the traffic generated by backbone i's websites brings reception revenue for the share α_i of the traffic that remains on-net.

[13] If consumers do not derive any utility from receiving calls ($v = 0$), as in Laffont, Rey, and Tirole (1998a), the price p_i cannot be positive; networks could, however, subsidize receivers.

[14] In practice, however, websites may gain from enhanced reliability or redundancy, at the cost of diseconomies of scale in the interface with the backones.

[15] For example, European or Australian Internet service providers must be connected to U.S. backbones through costly transoceanic cables that raise both origination and termination costs relative to a U.S.-based customer.

© RAND 2003.

In practice, this cost-based price discrimination may be implemented by setting different charges for local delivery; alternatively, it can be implemented by uniform charges applied at given points of interconnection, together with the requirement of the provision by the end users (or their ISPs) of circuits leading to these points of interconnection.

(vi) *Installed bases*. Suppose that backbone i has an installed base $\hat{\alpha}_i$ of consumers and an installed base $\hat{\tilde{\alpha}}_i$ of websites that are, for example, engaged in long-term contracts. Let $\hat{p}_i$ and $\hat{\tilde{p}}_i$ denote the predetermined prices charged to installed base consumers and websites by network i. The operators' profits become

$$\pi_i = \alpha_i\,[p_i - (c_t - a)] + \tilde{\alpha}_i\,[\tilde{p}_i - (c_o + a)] + \hat{\alpha}_i\,[\hat{p}_i - (c_t - a)] + \hat{\tilde{\alpha}}_i[\hat{\tilde{p}}_i - (c_o + a)].$$

Consequently, the equilibrium prices are unchanged: new customers are charged the off-net-cost prices and operator i's equilibrium profit $\pi_i^*(a)$ is equal to

$$\frac{d\pi_i^*}{da} = \hat{\alpha}_i - \hat{\tilde{\alpha}}_i.$$

Two simple implications can be drawn from this observation. First, web-hosting backbones prefer a low termination charge, while backbones that are stronger on the dial-up side, say, prefer a high termination charge. Second, if the termination charge is determined in a private negotiation, two backbones tend to have conflicting interests if one leans much more heavily to one side of the market than does the other. However, their interests do not necessarily conflict (even if one carries far more traffic than the other) if, say, one segment of the market has (for both backbones) developed more quickly than the other segment.

4. Ramsey access charges

■ By focusing on inelastic demands, the benchmark model of Section 2 and the various extensions performed in Section 3 sidestepped welfare issues. This section maintains the perfect-competition assumption but allows for elastic demands. Perfect competition implies that backbones' budgets are always balanced, whatever the access charge. But through its allocation of costs between end users, the access charge plays a central role in achieving economic efficiency. We show below that the Ramsey access charges, i.e., the access charges that maximize social welfare, must take into account not only the demand elasticities of the two segments, but also the externality that each side exerts on the other.[16]

Suppose for example that a consumer derives surplus v, drawn from a distribution $F(v)$, from being connected with websites; similarly, a website derives a surplus $\tilde{v}$, drawn from a distribution $\tilde{F}(\tilde{v})$, from being connected with consumers. Consumers' and websites' demands are thus given by $q = D(p) = 1 - F(p)$ and $\tilde{q} = \tilde{D}(\tilde{p}) = 1 - \tilde{F}(\tilde{p})$.[17] Furthermore, consumers' and websites' *net* surpluses are given by $S(p) = \int_p^{+\infty}(v - p)dF(v)$ and $\tilde{S}(\tilde{p}) = \int_{\tilde{p}}^{+\infty}(\tilde{v} - \tilde{p})d\tilde{F}(\tilde{v})$. Then,

[16] Similar conclusions hold for the credit card industry, in which the "backbones" are "banks," the "websites" and "consumers" are the "merchants" and "cardholders," and the "access charge" is the "interchange fee." See Rochet and Tirole (2002), Schmalensee (2002), Schwartz and Vincent (2000), and Wright (2000). Related insights apply to B2B—see Caillaud and Jullien (2001).

[17] There again, prices can be interpreted as pure traffic-based prices or as subscription prices (per website reached or per consumer reached). For example, if backbones simply charge a subscription price T for receiving traffic, the relevant consumer price is $p = T/\tilde{D}$, where $\tilde{D}$ denotes the number of connected websites, and a consumer with a valuation v subscribes again if $p \leq v$.

© RAND 2003.

Proposition 2. When both consumers' and websites' demands are elastic, the Lindahl (first-best) prices are given by

$$p^{FB} + \tilde{p}^{FB} = c - \frac{S\left(p^{FB}\right)}{D\left(p^{FB}\right)} = c - \frac{\tilde{S}\left(\tilde{p}^{FB}\right)}{\tilde{D}\left(\tilde{p}^{FB}\right)},$$

whereas the Ramsey (second-best) prices and access charge are characterized by $p^{SB} = c_t - a^{SB}$, $\tilde{p}^{SB} = c_o + a^{SB}$, and

$$\frac{S\left(p^{SB}\right)}{D'\left(p^{SB}\right)} = \frac{\tilde{S}\left(\tilde{p}^{SB}\right)}{\tilde{D}'\left(\tilde{p}^{SB}\right)}. \tag{4}$$

Proof. Social welfare is equal to

$$W = S(p)\tilde{D}(\tilde{p}) + D(p)\tilde{S}(\tilde{p}) + (p + \tilde{p} - c)D(p)\tilde{D}(\tilde{p});$$

its first-best level is thus characterized by

$$p + \tilde{p} = c - \frac{S(p)}{D(p)} = c - \frac{\tilde{S}(\tilde{p})}{\tilde{D}(\tilde{p})}.$$

Ramsey prices maximize W subject to the budget constraint

$$(p + \tilde{p} - c)D(p)\tilde{D}(\tilde{p}) \geq 0.$$

Denoting by λ the Lagrangian multiplier associated with this budget constraint, and using $p+\tilde{p} = c$, the first-order conditions boil down to

$$-\lambda = D'(p)\tilde{S}(\tilde{p}) = \tilde{D}'(\tilde{p})S(p).$$

Q.E.D.

From a first-best perspective, each segment is charged a price equal to the marginal cost, minus a discount that reflects the positive externality exerted on the other segment. For example, an extra website generates an additional gross consumer surplus $S + pD$, so that the (per-consumer) price $\tilde{p}$ charged to websites must be decreased by an amount equal to the (per-capita, or average) consumer surplus $v^e \equiv p + S/D$:

$$\tilde{p} = c - v^e.$$

Similarly, the (per-website) price charged to consumers must be discounted for the average surplus $\tilde{v}^e \equiv \tilde{p} + \tilde{S}/\tilde{D}$ that consumers bring to websites: $p = c - \tilde{v}^e$. Since average surpluses exceed prices ($v^e > p$, $\tilde{v}^e > \tilde{p}$), the total price charged to the two segments, $p + \tilde{p}$, must be lower than the cost c; the subsidy must reflect the positive externality that each segment exerts on the other:

$$c - (p + \tilde{p}) = \frac{S(p)}{D(p)} = \frac{\tilde{S}(\tilde{p})}{\tilde{D}(\tilde{p})},$$

which in particular implies that, at the optimum, these two externalities must be equalized.

In a second-best world, the budget constraint rules out outside subsidies. Prices must therefore be increased so as to fully cover the cost c, according to standard Ramsey principles: the departure from first-best prices should be inversely related to the magnitude of demand elasticities:

$$\frac{p - (c - \tilde{v}^e)}{p} = \frac{\lambda}{\eta}, \quad \frac{\tilde{p} - (c - v^e)}{\tilde{p}} = \frac{\lambda}{\tilde{\eta}},$$

© RAND 2003.

where η and $\tilde{\eta}$ denote the demand elasticities and λ the Lagrangian multiplier associated with the budget constraint. In the absence of fixed cost, the budget constraint is simply that the total price $p + \tilde{p}$ must cover the joint cost c and the above Ramsey formulas boil down to (4), which can be interpreted as follows. Increasing the consumer price discourages some consumers, which reduces website surplus; the corresponding welfare loss is thus given by $D'(p)\tilde{S}(\tilde{p})$. Similarly, increasing the website price discourages some websites, which reduces consumer surplus, thereby generating a welfare loss $\tilde{D}'(\tilde{p})S(p)$. The optimal tradeoff thus depends on how many end users are discouraged on one side, as well as on the net surplus lost on the other side, and balances the two types of welfare losses: $D'(p)\tilde{S}(\tilde{p}) = \tilde{D}'(\tilde{p})S(p)$. A special case occurs when one side of the market is inelastic as in Section 2; then, the access charge shifts the burden as much as possible to the inelastic segment.

Remark. In sharp contrast with the recommendations usually derived from standard Ramsey pricing formulas, the tradeoff just described can lead, in equilibrium, to a higher price for the segment with the higher elasticity. To see this, note that condition (4) can be rewritten as (letting $\eta_S = -pS'/S$ and $\tilde{\eta}_{\tilde{S}} = -\tilde{p}\tilde{S}'/\tilde{S}$)

$$\left(\frac{p}{\tilde{p}}\right)^2 = \frac{\eta}{\tilde{\eta}}\frac{\eta_S}{\tilde{\eta}_{\tilde{S}}}.$$

That is, prices in the two segments should covary with their respective demand elasticities (η or $\tilde{\eta}$) (and with the related surplus elasticities, η_S and $\tilde{\eta}_{\tilde{S}}$).

Under perfect competition, firms make zero profit; they are thus indifferent as to the level of the access charge and should not resist a regulation of the access charge that implements the second-best optimum. In practice, backbones have historically opted for "bill and keep" ($a = 0$), which minimizes transaction costs. Bill and keep is favorable to websites,[18] which might have been a good idea to promote the development of Internet-based services. Now that many web services are available, and the emphasis is more on encouraging consumers to connect and use these services, absent significant transaction costs, bill and keep is unlikely to be close to optimal.

5. Amending the off-net-cost pricing principle

■ **Variable demand and two-part tariffs.** Let us extend the model to allow for variable demand functions and connection costs for consumers—sticking to the same formulation as before for websites. It is then natural to also allow backbones to charge two-part tariffs to consumers. Because of the connection costs, the off-net costs no longer predict average retail prices; however, they still define the relevant marginal usage prices if backbones compete in nonlinear tariffs. To see this, for $i = 1, 2$, let p_i denote the volume-based fee and F_i the fixed fee charged by backbone i, and $D(p_i)$ the demand of a representative consumer who subscribes, with $S(p_i)$ the associated net surplus (but gross of the fixed fee F_i). A consumer thus subscribes to backbone i if

$$S(p_i) - F_i > S(p_j) - F_j.$$

Backbone i's profit is then given by

$$\begin{aligned}\pi_i &= \alpha_i (F_i - f) + \alpha_i\tilde{\alpha}_i D(p_i)(p_i + \tilde{p}_i - c) + \alpha_i\tilde{\alpha}_j D(p_j)(p_i - c_t + a)\\ &\quad + \alpha_j\tilde{\alpha}_i D(p_j)(\tilde{p}_i - c_o - a)\\ &= \alpha_i (F_i - f) + \alpha_i(\tilde{\alpha}_1 + \tilde{\alpha}_2) D(p_i)[p_i - (c_t - a)]\\ &\quad + \tilde{\alpha}_i[\alpha_1 D(p_1) + \alpha_2 D(p_2)][\tilde{p}_i - (c_o + a)].\end{aligned}$$

[18] When $a = 0$, consumers pay the entire termination cost, which, as noted above, is in practice the larger part of the cost due to "hot-potato" routing.

© RAND 2003.

The opportunity cost of stealing a website away from the rival network is

$$\alpha_i D(p_i)[c - (c_t - a)] + \alpha_j D(p_j)[(c_o + a) - 0] = (c_o + a)q,$$

where $q = \alpha_1 D(p_1) + \alpha_2 D(p_2)$ denotes the volume of traffic generated by each website. The opportunity cost of stealing a website, per unit of traffic, is thus again $c_o + a$; therefore, in equilibrium, $\tilde{p}_1 = \tilde{p}_2 = c_o + a$.

Also, *if* $p_1 = p_2 = p$, then the opportunity cost of stealing a consumer away from the rival network is

$$\tilde{\alpha}_i D(p)[c - (c_o + a)] + \tilde{\alpha}_j D(p)[(c_t - a) - 0] + f - F_i = (c_t - a) D(p) + f - F_i;$$

furthermore, *if* $\tilde{p}_i = c_o + a$, then the opportunity cost of inducing its own consumers to generate one more unit of traffic is similarly given by

$$\tilde{\alpha}_i D(p)[c - \tilde{p}_i] + \tilde{\alpha}_j D(p)[(c_t - a) - 0] = (c_t - a) D(p).$$

Therefore, the off-net-cost pricing principle still applies, although now *only* in equilibrium. We thus have the following:

Proposition 3. When $\tilde{v} \geq c_o + a$ and $S(c_t - a) \geq f$, there exists a unique two-part-tariff equilibrium, given by

$$\begin{aligned} p_1 &= p_2 = p^* = c_t - a, \\ \tilde{p}_1 &= \tilde{p}_2 = \tilde{p}^* = c_o + a, \\ F_1 &= F_2 = f, \\ \pi_1 &= \pi_2 = \pi^* = 0. \end{aligned}$$

The off-net-cost pricing principle therefore still applies: in equilibrium, the fixed fee is equal to the connection cost, and usage prices for sending and receiving traffic are equal to the off-net costs of outgoing and incoming traffic.

□ **Market power.** Section 3 has demonstrated the remarkable robustness of the off-net-cost pricing principle in a competitive industry. We now investigate how the principle must be amended if the backbones have some market power, which supposes that they provide differentiated services. Intuitively, the relevant marginal cost remains the off-net-cost, but a markup should be added because of market power. We will say that an access charge is "constrained Ramsey optimal" if it is optimal when the only regulatory instrument is the access charge (that is, when market power cannot be directly addressed).

Let us maintain the assumption that the backbones are perfectly substitutable on the consumer segment but introduce some differentiation on the website segment. Websites' surplus from being connected with consumers then depends on the network to which they subscribe.

Backbones can engineer their networks to provide different levels of quality, in several dimensions. They can, for example, ensure that capacity is significantly higher than offered load,[19] and can do so with a focus on either the mean or the variance of delay, which may be of some interest for different users.[20] Backbones can also invest in reliability.[21] Higher speed, lower

[19] By spending more to build out more capacity, but also by doing a better job of predicting demand or by using caching or replication to reduce the amount of traffic they haul around the network. The straightness of the fiber runs and the number of router hops can also play a role.

[20] For example, real-time voice requires a low average delay and low variability, whereas email users might not care much about variability.

[21] E.g., by ensuring redundancy. Simplicity can also contribute to reliability, as well as good connectivity to other networks.

© RAND 2003.

delay variability, and higher reliability have an impact on the benefit that a website can derive from connecting to alternative backbones, and this impact depends on the website's business model.

A different kind of engineering enhances reliability. Ensuring redundancy, with no single point of failure, can be important. Simplicity can also contribute to reliability, but paradoxically is often somewhat at odds with extensive redundancy. Good connectivity to other networks can also be a differentiator. This tends to play a larger role in periods where, for one reason or another, connectivity between networks is in general less than ideal.

These reliability and quality differences have an impact on the benefit that a website can derive from connecting to the alternative backbones, and this impact depends on the website's business model. Letting $\tilde{v}_1$ and $\tilde{v}_2$ denote the surpluses that a website derives when subscribing respectively to backbone 1 or 2, the website subscribes to network i if (for $i \neq j = 1, 2$)[22]

$$\tilde{v}_i - \tilde{p}_i \geq \max\left\{0, \tilde{v}_j - \tilde{p}_j\right\}.$$

The values $\tilde{v}_1$ and $\tilde{v}_2$ are distributed among websites according to $\tilde{F}(\tilde{v}_1, \tilde{v}_2)$ on $\Re^2_+$, which determines the number $\tilde{D}_i(\tilde{p}_i, \tilde{p}_j)$ of websites choosing to subscribe to network i; by construction, an increase in one operator's price increases the demand for the other operator but reduces the total number of connected websites: $\partial\tilde{D}_j/\partial\tilde{p}_i > 0$ and $\partial\tilde{D}_i/\partial\tilde{p}_i + \partial\tilde{D}_j/\partial\tilde{p}_i < 0$. We will furthermore maintain the following assumptions:

(i) $\tilde{D}_1(\tilde{p}_1, \tilde{p}_2) = \tilde{D}_2(\tilde{p}_2, \tilde{p}_1)$. Therefore $\tilde{D}(\tilde{p}) \equiv \tilde{D}_i(\tilde{p}, \tilde{p}) = [1 - \tilde{G}(\tilde{p})]/2$, where $\tilde{G}(\tilde{v})$ is the cumulative distribution of $\tilde{v} \equiv \max\{v_1, v_2\}$, and decreases when $\tilde{p}$ increases: $\tilde{D}' < 0$.

(ii) For every $\tilde{c}$ there exists a unique price $\tilde{p}(\tilde{c})$ such that $\tilde{p} = \arg\max_{\tilde{p}'}(\tilde{p}' - \tilde{c})\tilde{D}_i(\tilde{p}', \tilde{p})$ and $\tilde{\pi}(\tilde{c}) \equiv [\tilde{p}(\tilde{c}) - \tilde{c}]\tilde{D}[\tilde{p}(\tilde{c})]$ decreases when $\tilde{c}$ increases: $\tilde{\pi}' < 0$.

The price $\tilde{p}$ and profit $\tilde{\pi}$ can be interpreted as benchmark equilibrium price and (per-backbone) profit in a hypothetical economy where the two backbones face demands $\tilde{D}_1$ and $\tilde{D}_2$ and compete with the same cost $\tilde{c}$. The assumptions on the demand and profits are plausible: an increase in either backbone's price reduces the number of connected websites, and an increase in the industry marginal cost reduces equilibrium profits.[23]

Inelastic consumer demand. Suppose first that all consumers get the same value v from being connected to a website. Then, normalizing to 1 the population of consumers, we have the following:

Proposition 4. Assume that consumer demand is inelastic and $v \geq c_t - a$.[24] Then there exists a unique equilibrium. This equilibrium is symmetric and satisfies

$$p_i = p^* = c_t - a, \quad \tilde{p}_i = \tilde{p}\left(c_o + a\right), \quad \pi_i = \tilde{\pi}\left(c_o + a\right).$$

Increasing the access charge raises the equilibrium website price but reduces both the number of connected websites and the equilibrium profits. Backbones favor the lowest admissible access charge, $a_\Pi = c_t - v$, which fully extracts the surplus from consumers and subsidizes the profitable website segment; consumer demand being inelastic, this access charge is constrained Ramsey optimal.

[22] As before, we assume that each consumer gets the same surplus from connecting to every website and, similarly, that each website gets the same surplus from connecting to every consumer. This allows for a simpler analysis of consumers' and websites' subscription decisions—and in particular eliminates any interdependence between these decisions.

[23] This condition is, for example, always satisfied when backbones absorb part of the cost increase ($0 < \tilde{p}' < 1$), since then both the demand and the margin decrease when the cost increases.

[24] For $v < c_t - a$, there is no symmetric equilibrium in pure strategies—but an asymmetric equilibrium may exist, in which one backbone serves all consumers at a price $p = v$.

© RAND 2003.

Proof. Denoting by α_i backbone i's share of consumers, the profit of that backbone is given by

$$\pi_i = \alpha_i \left(\tilde{D}_1 + \tilde{D}_2\right) [p_i - (c_t - a)] + \tilde{D}_i \left[\tilde{p}_i - (c_o + a)\right].$$

The standard Bertrand argument thus applies to the inelastic consumer segment. As long as it remains below v, the price paid by consumers does not affect their demand and thus has no effect on the website business. Therefore, $p_1 = p_2 = c_t - a$, which by assumption does not exceed v, and network i's profit boils down to

$$\pi_i = \tilde{D}_i \left(\tilde{p}_i, \tilde{p}_j\right) [\tilde{p}_i - (c_o + a)],$$

which leads at a symmetric equilibrium to $\tilde{p}_i = \tilde{p}(c_o + a)$ and $\pi_i = \tilde{\pi}(c_o + a)$. *Q.E.D.*

Differentiation weakens the intensity of price competition for websites and allows backbones to earn a positive profit on that segment; website prices are thus higher than the off-net cost $c_o + a$, but prices and profits are "as if" backbones were competing for websites with a cost equal to their off-net cost. A lower access charge increases the opportunity cost of servicing consumers, leading the networks to raise the consumer price. On the other side, the reduction in the access charge lowers the opportunity cost of servicing the websites, which generates more profit on websites. In other words, decreasing the access charge allows the networks to extract more rents from the surplus that consumers derive from receiving traffic; part of those rents are passed on to websites, which benefit from a reduction in the price they are charged for sending traffic to consumers, and part of it serves to increase networks' profits.

Backbones thus favor the lowest access charge a, subject to consumers' participation constraint ($v \geq c_t - a$). Since consumer demand is inelastic, backbones' interest is here in line with the social interest, which also requires extracting consumer rents in order to subsidize and attract more websites.

Elastic consumer demand. Suppose now that consumers' surplus v is drawn from a distribution $F(v)$. Consumers' demand is thus $D(p) = 1 - F(p)$. The following assumption guarantees that profit functions are "well behaved" and rules out the possibility that a backbone might desire making losses on consumers in order to enhance profits on the website segment.

Assumption 1. There exists $k > 0$ such that

(i) $\forall \tilde{c}, \quad \tilde{p}(\tilde{c}) - \tilde{c} < k$;
(ii) $\forall p, \quad D(p) + kD'(p) \geq 0$.

Assumption 1 ensures that prices and profits are again "as if" backbones were (imperfectly) competing on each segment with a cost equal to their off-net cost:

Proposition 5. With an elastic consumer demand, there exists a unique symmetric equilibrium under Assumption 1. This equilibrium satisfies

$$p_i = p^* = c_t - a, \quad \tilde{p}_i = \tilde{p}(c_o + a), \qquad \pi_i = \tilde{\pi}(c_o + a) D(c_t - a).$$

An increase in the access charge raises the equilibrium website price and thus reduces the number of connected websites, but it decreases the equilibrium consumer price and thus attracts more consumers.

Proof. Fix $p_j = p = c_t - a$ and $\tilde{p}_j = \tilde{p} = \tilde{p}(c_o + a)$. By raising its price p_i above $p = c_t - a$, backbone i gets $[\tilde{p}_i - (c_o + a)]\tilde{D}_i(\tilde{p}_i, \tilde{p})D(p)$ and thus cannot earn more than (and actually earns exactly) $\tilde{\pi}(c_o + a)D(c_t - a)$. By reducing its price p_i below $c_t - a$, backbone i's profit is of the form $\pi_i(p_i, \tilde{p}_i) = D(p_i)\hat{\pi}_i(p_i, \tilde{p}_i)$, where

$$\hat{\pi}_i (p_i, \tilde{p}_i) = [p_i - (c_t - a)] \left[\tilde{D}_i (\tilde{p}_i, \tilde{p}) + \tilde{D}_j (\tilde{p}, \tilde{p}_i)\right] + [\tilde{p}_i - (c_o + a)] \tilde{D}_i (\tilde{p}_i, \tilde{p})$$

© RAND 2003.

is maximal for some $\tilde{p}_i(p_i) > \tilde{p}$: the loss on the consumer segment gives an incentive to reduce the traffic and thus the number of connected websites, $\tilde{D}_i + \tilde{D}_j$. Using the envelope theorem, the impact of p_i on backbone i's maximal profit is given by

$$\frac{d\pi_i\left(p_i, \tilde{p}_i\left(p_i\right)\right)}{dp_i} = D\left(p_i\right)\left[\tilde{D}_i\left(\tilde{p}_i\left(p_i\right), \tilde{p}\right) + \tilde{D}_j\left(\tilde{p}, \tilde{p}_i\left(p_i\right)\right)\right] + D'\left(p_i\right)\tilde{\pi}_i\left(p_i, \tilde{p}_i\left(p_i\right)\right)$$
$$\geq D\left(p_i\right)\tilde{D}\left(\tilde{p}\right) + D'\left(p_i\right)\left[\tilde{p} - \left(c_o + a\right)\right]\tilde{D}\left(\tilde{p}\right),$$

where the inequality stems from (using $\tilde{p}_i \geq \tilde{p}$)

$$\tilde{D}_i\left(\tilde{p}_i, \tilde{p}\right) + \tilde{D}_j\left(\tilde{p}, \tilde{p}_i\right) \geq \tilde{D}_j\left(\tilde{p}, \tilde{p}_i\right) \geq \tilde{D}_j\left(\tilde{p}, \tilde{p}\right) = \tilde{D}\left(\tilde{p}\right)$$

and (using $p_i \leq p$)

$$\tilde{\pi}_i\left(p_i, \tilde{p}_i\left(p_i\right)\right) \leq \tilde{\pi}_i\left(p, \tilde{p}_i\left(p\right)\right) = \left[\tilde{p} - \left(c_o + a\right)\right]\tilde{D}\left(\tilde{p}\right).$$

Assumption 1 then ensures that $d\pi_i(p_i, \tilde{p}_i(p_i))/dp_i \geq 0$ for all $p_i < c_t - a$, implying that backbones cannot gain from subsidizing consumers. *Q.E.D.*

The off-net-cost pricing principle still applies: in each segment, backbones' equilibrium prices are as if backbones' marginal cost were equal to the off-net cost; they are exactly equal to the off-net cost of receiving traffic in the competitive consumer segment and correspond to the oligopolistic price $\tilde{p}(c_o + a)$ in the website segment. Let us now assume that backbones partially pass through cost increases to websites:

Assumption 2. $0 < \tilde{p}' < 1$.

Under this reasonable assumption, by adjusting the access charge the backbones move both prices, p and $\tilde{p}$, in such a way that $d\tilde{p}/dp = -\tilde{p}'$ lies between -1 and 0.

Backbones' preferred access charge maximizes the per-backbone profit

$$\Pi = \left(p + \tilde{p} - c\right)D\left(p\right)\tilde{D}\left(\tilde{p}\right).$$

The privately optimal access charge trades off the impact on the two prices p and $\tilde{p}$ and satisfies

$$\Pi' = \left(1 + \frac{d\tilde{p}}{dp}\right)D\left(p\right)\tilde{D}\left(\tilde{p}\right) + \left(p + \tilde{p} - c\right)\left[D'\left(p\right)\tilde{D}\left(\tilde{p}\right) + D\left(p\right)\tilde{D}'\left(\tilde{p}\right)\frac{d\tilde{p}}{dp}\right] = 0.$$

Given the partial pass-through assumption, $1 + \frac{d\tilde{p}}{dp} > 0$, moving some of the communication costs from the website segment to the consumer segment increases the competitive consumer price more than it reduces the less competitive website price—in that segment, backbones keep part of the reduction in the form of increased margins. Therefore, the first term is negative and gives backbones an incentive to use the consumer segment for subsidizing the website segment. The other term reflects a profit motivation for a large volume of communication, and it increases in magnitude, the larger the total margin $p + \tilde{p} - c$.

Backbones' preferred access charge can be compared with the constrained Ramsey optimal one, which maximizes total welfare given market power on the website segment. The latter is given by

$$\max_{\{p,\tilde{p}\}} W = \left(v^e + \tilde{v}^e - c\right)D\left(p\right)\tilde{D}\left(\tilde{p}\right),$$

where v^e and $\tilde{v}^e$ represent consumers' and websites' average surplus:

$$v^e = \frac{\int_p^{+\infty} v dF\left(v\right)}{D\left(p\right)}, \quad \tilde{v}^e = \frac{\int_p^{+\infty} \tilde{v} d\tilde{G}\left(\tilde{v}\right)}{\tilde{D}\left(\tilde{p}\right)}.$$

© RAND 2003.

We thus have[25]

$$W' = \left(p + \tilde{v}^e - c\right) D'(p)\, \tilde{D}(\tilde{p}) + \left(v^e + \tilde{p} - c\right) D(p)\, \tilde{D}'(\tilde{p}) \frac{d\tilde{p}}{dp}$$
$$= \Pi' + \left(1 + \frac{d\tilde{p}}{dp}\right) D(p)\, \tilde{D}(\tilde{p}) + \left(\tilde{v}^e - \tilde{p}\right) D'(p)\, \tilde{D}(\tilde{p}) + \left(v^e - p\right) D(p)\, \tilde{D}'(\tilde{p}) \frac{d\tilde{p}}{dp}.$$

As compared with backbones' preferred access charge, for which $\Pi' = 0$, the first additional term corresponds to a social gain from lowering consumer prices at the expense of websites. This stems from the fact that, as noted above, backbones have an excessive incentive to shift communication costs toward the more competitive consumer segment, since this cost is then passed through to consumers; backbones then benefit from lower costs in the less competitive website segment, where they pocket part of the cost reduction. The other two additional terms derive from the fact that backbones do not fully appropriate their customers' surplus. Thus, for example, attracting one more consumer gives websites an additional surplus $(\tilde{v}^e - \tilde{p})\tilde{D}(\tilde{p})$ that backbones cannot grab.

Note that, for the access charge a_Π that maximizes profits, we have

$$W' = \left(1 + \frac{\tilde{v}^e - \tilde{p}}{p + \tilde{p} - c}\right)\left(1 + \frac{d\tilde{p}}{dp}\right) D(p)\, \tilde{D}(\tilde{p}) + \left[\left(v^e - p\right) - \left(\tilde{v}^e - \tilde{p}\right)\right] D(p)\, \tilde{D}'(\tilde{p}) \frac{d\tilde{p}}{dp}.$$

Therefore we have the following:

Proposition 6. With an elastic consumer demand and under Assumptions 1 and 2, backbones prefer an access charge that is lower than the socially optimal one, thereby favoring websites, whenever either (i) backbones leave more surplus to customers on the consumer segment ($v^e - p > \tilde{v}^e - \tilde{p}$ for $a = a_\Pi$) or (ii) backbones appropriate most of the cost reduction on the website segment ($d\tilde{p}/dp$ close to zero).

□ **Asymmetric access charges.** While the basic insight of the benchmark model has very broad applicability, the symmetric-access-charge assumption is crucial. We now demonstrate that asymmetric access charges are a factor of instability. Consider first the competitive backbone industry of Section 2, and now let a_i denote the access charge paid by backbone $j \neq i$ to backbone i for terminating backbone j's off-net traffic. Without loss of generality, let us assume that

$$a_1 > a_2.$$

A first intuition is that the high-access-charge backbone 1 has a comparative advantage for both consumers (since receiving traffic is particularly attractive to this network) and websites (since terminating traffic on the rival backbone is cheaper for backbone 1). This reasoning, however, fails to account for opportunity costs. For example, if network 1 makes much money when its consumers download from network 2's websites, for the same reason network 2 finds it costly to leave consumers to network 1.

A second observation is that backbone 1 has an incentive to focus on one side of the market so as to generate off-net traffic, whereas backbone 2 has an incentive to be present on both markets so as to avoid off-net traffic. To see this, note that backbone i's profit can be written as

$$\pi_i = \alpha_i \tilde{\alpha}_j \left(p_i + \tilde{p}_i - c\right) + \alpha_i \tilde{\alpha}_j \left[p_i - (c_t - a_i)\right] + \alpha_j \tilde{\alpha}_i \left[\tilde{p}_i - \left(c_o + a_j\right)\right]$$
$$= \alpha_i \left[p_i - (c_t - a_i)\right] + \tilde{\alpha}_i \left[\tilde{p}_i - \left(c_o + a_j\right)\right] + \alpha_i \tilde{\alpha}_i \left(a_j - a_i\right), \tag{5}$$

if all potential end users are connected.

[25] See Schmalensee (2002) for a similar derivation in the context of the credit card industry.

© RAND 2003.

Backbone 2's gain from a simultaneous increase in both α_2 and $\tilde{\alpha}_2$ exceeds the sum of the gains obtained by increasing α_2 or $\tilde{\alpha}_2$ alone. Similarly, backbone 1 gains more when it simultaneously increases its market share on one side and reduces its market share on the other. As a result of these conflicting interests, there is no equilibrium in pure strategies:

Proposition 7. If the backbones charge asymmetric access charges, then there is no pure-strategy equilibrium.

Proof. See the Appendix.

This inexistence problem stems from the fact that, as just noted, asymmetric access charges make the profits nonconcave.[26] This nonconcavity problem is somewhat robust. Suppose, for example, that the backbones are horizontally differentiated à la Hotelling on both segments, with differentiation parameters t and $\tilde{t}$ and a unit length on each segment. The profit functions are still as in (5), but market shares are now given by

$$\alpha_i = \frac{1}{2} + \frac{1}{2t}\left(p_j - p_i\right),$$
$$\tilde{\alpha}_i = \frac{1}{2} + \frac{1}{2\tilde{t}}\left(\tilde{p}_j - \tilde{p}_i\right).$$

Therefore,

$$\frac{\partial^2 \pi_i}{\partial p_i^2} = -\frac{1}{t}, \ \frac{\partial^2 \pi_i}{\partial \tilde{p}_i^2} = -\frac{1}{\tilde{t}}, \ \frac{\partial^2 \pi_i}{\partial p_i \partial \tilde{p}_i} = \frac{a_j - a_i}{t\tilde{t}}$$

and thus profits are concave only if

$$t\tilde{t} \geq (a_1 - a_2)^2,$$

that is, backbones must be sufficiently differentiated (and thus not competing too effectively) on both segments—and the required level of differentiation is proportional to the asymmetry in the access charges.[27]

Proposition 7 seems to call for reciprocal access charges. Reciprocity, however, should be understood in a broad sense, allowing for termination cost differences. For example, suppose that backbone 1 has a more expensive "shortest-exit" policy; backbone 1 then bears a larger proportion of termination transportation cost on off-net traffic: $c_t^1 = c_t^2 + \Delta$ (and thus $c_o^2 = c_o^1 - \Delta$). Then a (pure-strategy) equilibrium exists only when backbones account for this cost asymmetry when setting their access charges, that is, $a_1 = a_2 + \Delta$ (the competitive prices are then $p^* = c_t^1 - a_1 = c_t^2 - a_2$ and $\tilde{p}^* = c_o^1 + a_2 = c_o^2 + a_1$). That is, the backbone that keeps off-net traffic on its own network longer before delivering it to the other should be "rewarded" by being charged a lower termination fee.

6. Micropayments and neutrality

■ An increase in the access charge raises the cost for websites of doing business. Websites then may be tempted to pass through the increased traffic-related cost to the consumers who request the traffic. With some exceptions, such traffic-based "micropayments" do not yet exist. They require putting in place costly billing and end-user micropayment information systems.

[26] This nonconcavity—and the issue of the existence of an equilibrium—is reminiscent of the analysis of two-way interconnection between telecom operators, as in Laffont, Rey, and Tirole (1998a). Due to the "missing price," however, in that case this problem appears even with symmetric access charges.

[27] The characterization of an access charge equilibrium in pure strategies would thus be achievable when backbones are sufficiently differentiated.

© RAND 2003.

If websites pass their cost of traffic through to the consumers, the consumers' final demand does not depend on the share of termination cost that they pay directly (through the price p for receiving traffic) but rather on the total price of the communication $(p + \widetilde{p})$. This in turn suggests that the way in which the total cost is *a priori* distributed between senders and receivers is irrelevant. Put differently, the access charge, which mainly affects how the cost of traffic is divided between senders and receivers, may have no impact on the consumers' final demand and thus on traffic volume. This section shows that in many contexts, the access charge is indeed neutral, i.e., it has no impact on traffic and efficiency.

We consider four illustrations. In the first, there is perfect competition at both the backbones' and websites' levels, and websites can charge consumers (through "micropayments") for their cost of traffic. As a result, backbones charge senders and receivers according to their perceived opportunity costs, as before, but consumers end up incurring the total cost of traffic regardless of the level of the access charge. The next two illustrations show that the access charge remains neutral when the consumers have an elastic demand for websites' services and when websites are not perfectly competitive. The last illustration considers situations where consumers use websites to buy goods or services. To the extent that the amount of communication is related to the volume of transactions on the goods and services, the price charged for those goods and services can play the role of micropayments. In the case of perfect correlation between communications and transactions, the access charge is again neutral, even if backbones do not perfectly compete for websites.

□ **Perfect competition at the backbone and website levels.** Let us assume that micropayments are feasible and costless. The pricing behavior of the websites depends on the degree of competition between them. Let us start with the case where there are multiple identical websites of each "type."We otherwise assume that the industry is as described in Section 2; in particular, backbones are perfect competitors on both sides, and consumers want to download one unit of traffic from each type of website. The timing goes as follows. After agreeing on an access charge, the backbones set prices (p_i for consumers, $\tilde{p}_i$ for websites). Then, the websites subscribe and choose micropayments (denoted by s) per unit of downloaded volume. Finally, the consumers subscribe and select websites.

The backbones' profits can still be written as

$$\pi_i = \alpha_i[p_i - (c_t - a)] + \tilde{\alpha}_i[\tilde{p}_i - (c_o + a)],$$

where, as before, for each category of end user (consumer or website), the market shares only depend on the prices charged to that category. This is clear for websites, which, by choosing the backbone with the lowest website price, not only minimize the cost of their traffic, but also enhance their competitive situation. But this is also true for consumers: given the micropayment s charged by a website, they face a total price $p_i + s$ if they subscribe to backbone i; they thus choose the backbone with the lowest consumer price.[28] As a result, off-net-cost pricing still prevails:

$$p_i = c_t - a,$$

and

$$\tilde{p}_i = c_o + a.$$

Bertrand websites set micropayments equal to their marginal net cost, which consists of their traffic cost, $\tilde{p}_i$, decreased by the value $\tilde{v}$ that they derive from consumers' visits. So websites

[28] With network-based price discrimination, the subscription decision of one category affects the price paid by the other category; in particular, websites would care about consumers' subscription decisions, since it would affect their competitive situation. Different timings with respect to subscription decisions may then lead to different coordination patterns.

© RAND 2003.

located on backbone i charge[29]

$$s_i = \tilde{p}_i - \tilde{v} = c_o + a - \tilde{v}.$$

This implies that consumers bear the full cost of web traffic net of the website's surplus, and that the access charge is neutral as regards the total price paid by consumers.[30] For all i,

$$p_i + s_i = c - \tilde{v}.$$

□ **Elastic demand for websites' services.** This neutrality result extends to the case where consumers have an elastic demand for websites' services, of the form $q = D(p + s)$. In that case, each category of end user still selects the backbone with the lowest price for that category, so that the volume of traffic is $\hat{D} = D(\min\{p_1, p_2\} + \min\{\tilde{p}_1, \tilde{p}_2\} - \tilde{v})$. Thus, backbones' profits can still be written as

$$\pi_i = \alpha_i \hat{D}[p_i - (c_t - a)] + \tilde{\alpha}_i \hat{D}[\tilde{p}_i - (c_o + a)], \tag{6}$$

and the standard Bertrand argument still applies to both categories of end users, so that again, $p = c_t - a$ and $\tilde{p} = c_o + a$. Hence, the volume of traffic is efficient: it is given by

$$D(p + \tilde{p} - \tilde{v}) = D(c - \tilde{v})$$

and is thus independent of the access charge.

□ **Imperfect competition among websites.** The neutrality result remains valid even when websites have market power. Suppose, for example, that there is only one website of each "type," therefore enjoying a monopoly position for this type. For notational simplicity, suppose also that $\tilde{v} = 0$; that is, the website does not derive any direct reputational or commercial benefit from the visit. As before, each category of end user selects the lowest price offered to that category, so that the relevant prices are $\tilde{p} = \min\{\tilde{p}_1, \tilde{p}_2\}$ and $p = \min\{p_1, p_2\}$. Given those prices, each website will choose s so as to maximize its profit, given by

$$(s - \tilde{p})\, D(p + s).$$

This amounts to choosing a "consumer price" $\hat{s} = p + s$ that maximizes $(\hat{s} - p - \tilde{p})\, D(\hat{s})$ and thus leads to

$$s = s^M(p + \tilde{p}) - p,$$

where

$$s^M(x) = \arg\max_s (s - x)\, D(s),$$

thereby generating a traffic

$$\hat{D} = D\left(s^M(p + \tilde{p})\right). \tag{7}$$

Therefore, backbones' profits are still given by (6), with $\hat{D}$ now given by (7). As a result, Bertrand competition between the two backbones leads again to off-net-cost pricing, $p_i = c_t - a$

[29] Since websites pass their traffic cost through to consumers, we need not make any assumption on $c_0 + a$ and $\tilde{v}$. As a result, s_i can be either positive or negative, depending on the value of the access charge a.

[30] A similar result can be found in Rochet and Tirole (2002) and Gans and King (2003) for the credit card industry. They provide conditions under which the removal of the no-discrimination rule (the rule forcing or inducing merchants (depending on the country) to charge the same price for cash and card payments) leads to a neutrality of the interchange fee.

© RAND 2003.

and $\tilde{p}_i = c_o + a$; the volume of traffic and each website's profit are given by $D(s^M(c))$ and

$$\pi_w^M = \left(s^M(c) - c\right) D(s^M(c)),$$

respectively, and are thus independent of the access charge. In addition, because of the websites' market power, consumers pay more than the cost of the web traffic, but the price that they face is not affected by the access charge.

□ **Websites selling goods and services.** This setup is also relevant when there is a transaction associated with the consumer visiting the site (e.g., Amazon.com selling books through its websites). If there is perfect correlation between the bandwidth usage and the size of the transaction, the price of the transaction can play a role similar to the micropayment s. Below we consider the case where consumers buy a commodity at unit price P after having browsed the website. Buying from the website involves two types of communication costs: a search cost (browsing, listening to samples of music, etc.), which websites usually do not charge to consumers, and downloading costs, which websites can recover through the price P of the commodity. We will focus here on the latter cost and assume that downloading requires bandwidth usage q. Accordingly, the demand function for the commodity is $D(P + pq)$.

Denoting by C the unit cost of production of the commodity, the websites' profit is given by

$$[P - C - \tilde{p}q]\, D(P + pq) = \left[\hat{P} - C - (p + \tilde{p})q\right] D\left(\hat{P}\right),$$

where $\hat{P} = P + pq$. From this expression we see that the optimal price $\hat{P}$, and consequently the demand for the commodity and thus the downloading traffic, depends only on the total price $p + \tilde{p}$. With perfect competition between backbones, this total price equals c, and therefore the equilibrium traffic is independent of the access charge. With imperfect competition of the type modelled in Section 5, this total price is higher than c but remains independent of the access charge, and so is the equilibrium traffic.

7. Summing up

■ We have developed a framework for Internet backbone competition, which has allowed us to analyze the impact of access charges on backbones' competitive strategies. As we have seen, in a broad range of environments the operators set prices for their customers *as if* the customers' traffic were entirely off-net. This comes from the fact that the opportunity cost of stealing traffic away from rival operators is indeed equal to the off-net cost of traffic. In addition, the opportunity cost of creating outgoing (incoming) traffic is again equal to the off-net cost of that traffic, provided that the price for receiving (sending) traffic itself reflects its own off-net cost.

Given this off-net-cost pricing principle, in the absence of direct payments between websites and consumers, the access charge determines the allocation of communication costs between senders (mainly websites) and receivers (mainly consumers) and thus affects the level of traffic. The socially optimal access charge takes into account the demand elasticities on the two segments, but also the magnitude of the externality that each segment generates on the other segment. Since perfectly competitive backbones are indifferent to the level of the access charge, they would not object to a regulation of this access charge. In contrast, if they have market power, backbone operators' interests are in general no longer aligned with social welfare, although assessing the bias in their ideal access charge requires detailed knowledge not only of the elasticities of demand and externalities, but also of the operators' relative market power vis-à-vis websites and consumers. Finally, when websites charge micropayments, or when websites sell goods and services, the impact of the access charge on welfare is reduced and is even neutral if websites can perfectly pass through the cost of traffic to their consumers.

© RAND 2003.

Appendix

■ Proof of Proposition 7. In any Bertrand equilibrium, the two backbones must charge the same prices (otherwise, the backbone charging the lower price could profitably raise that price):

$$p_1 = p_2 = p,$$
$$\tilde{p}_1 = \tilde{p}_2 = \tilde{p}.$$

Furthermore, since it could attract all end users by slightly undercutting both p and $\tilde{p}$, backbone 2 must get at least

$$\pi_2 \geq p + \tilde{p} - c.$$

Similarly, since backbone 1 can decide to attract all consumers or all websites, it must get at least

$$\pi_1 \geq \max \{p - (c_t - a_1), \tilde{p} - (c_o + a_2)\}.$$

However, the two backbones' joint profits cannot exceed $p + \tilde{p} - c$. Hence,

$$p + \tilde{p} - c \geq p + \tilde{p} - c + \max \{p - (c_t - a_1), \tilde{p} - (c_o + a_2)\},$$

which implies

$$p \leq c_t - a_1,$$
$$\tilde{p} \leq c_o + a_2.$$

Backbone 1's profit thus satisfies

$$\begin{aligned} \pi_1 &= \alpha_1 [p - (c_t - a_1)] + \tilde{\alpha}_1 [\tilde{p} - (c_o + a_2)] - \alpha_1 \tilde{\alpha}_1 (a_1 - a_2) \\ &\leq -\alpha_1 \tilde{\alpha}_1 (a_1 - a_2), \end{aligned}$$

and it can be nonnegative only if $\alpha_1 \tilde{\alpha}_1 = 0$, that is, if backbone 2 attracts all end users on at least one side of the market. Backbone 2's profit similarly satisfies

$$\begin{aligned} \pi_2 &= \alpha_2 [p - (c_t - a_2)] + \tilde{\alpha}_2 [\tilde{p} - (c_o + a_1)] + \alpha_2 \tilde{\alpha}_2 (a_1 - a_2) \\ &\leq (\alpha_2 \tilde{\alpha}_2 - \alpha_2 - \tilde{\alpha}_2)(a_1 - a_2) \end{aligned}$$

and is thus nonnegative only if $\alpha_2 = \tilde{\alpha}_2 = 0$, a contradiction. *Q.E.D.*

References

ARMSTRONG, M. "Network Interconnection in Telecommunications." *Economic Journal*, Vol. 108 (1998), pp. 545–564.

CAILLAUD, B. AND JULLIEN, B. "Competing Cybermediaries." *European Economic Review*, Vol. 45 (2001), pp. 797–808.

CARTER, M. AND WRIGHT, J. "Local and Long-Distance Network Competition." Mimeo, Universities of Canterbury and Auckland, 1999a.

——— AND ———. "Interconnection in Network Industries." *Review of Industrial Organization*, Vol. 14 (1999b), pp. 1–25.

CHERDRON. M. "Interconnection, Termination-Based Price Discrimination, and Network Competition in a Mature Telecommunications Market." Mimeo, Mannheim University, 2000.

DESSEIN, W. "Network Competition in Nonlinear Pricing." *RAND Journal of Economics*, forthcoming.

GANS, J. AND KING, S.P. "Using 'Bill-and-Keep' Interconnect Arrangements to Soften Network Competition." *Economics Letters*, Vol. 71 (2002) ,pp. 413–420.

——— AND ———. "The Neutrality of Interchange Fees in Payment Systems." *Topics in Economic Analysis and Policy*, Vol. 3 (2003), available at http://www.bepress.com/bejeap/topics/vol3/iss1/art1/.

GAO, L. "On Inferring Autonomous System Relationships in the Internet." Mimeo, 2000.

HAHN, J.H. "Network Competition and Interconnection with Heterogenous Subscribers." Mimeo, Oxford University, 2000.

HERMALIN, B.E. AND KATZ, M.L. "Network Interconnection with Two-Sided User Benefits." Working Paper, Haas School of Business, University of California at Berkeley, 2001.

© RAND 2003.

LAFFONT, J.-J., REY, P., AND TIROLE, J. "Network Competition: I. Overview and Nondiscriminatory Pricing." *RAND Journal of Economics*, Vol. 29 (1998a), pp. 1–37.

———, ———, AND ———. "Network Competition: II. Price Discrimination." *RAND Journal of Economics*, Vol. 29 (1998b), pp. 38–56.

MARCUS, S. *Designing Wide Area Networks and Internetworks: A Practical Guide.* Reading, Mass.: Addison-Wesley, 1999.

NRIC. "Service Provider Interconnection for Internet Protocol Best Effort Service." White paper of the Network Reliability and Interoperability Council, http://www.nric.org/fg/fg4/ISP_Interconnection.doc, 2002.

ROCHET, J.-C. AND TIROLE, J. "Cooperation Among Competitors: The Economics of Payment Card Associations." *RAND Journal of Economics*, Vol. 33 (2002), pp. 1–22.

SCHMALENSEE, R. "Payment Systems and Interchange Fees." *Journal of Industrial Economics*, Vol. 50 (2002), pp. 103–122.

SCHWARTZ, M. AND VINCENT, D.R. "The No Surcharge Rule in Electronic Payments Markets: A Mitigation of Pricing Distortions?" Mimeo, 2000.

WRIGHT, J. "An Economic Analysis of a Card Payment Network." Mimeo, University of Auckland, 2000.

© RAND 2003.

Exhibit E

Call Termination Fees: The U.S. in global perspective

J. Scott Marcus[1]

Abstract

The economic framework under which the United States implements call termination fees is unusual. Several recent studies suggest that the United States system has resulted in greater use of mobile telephony services and in lower cost to consumers than many other systems. This paper summarizes call termination fee mechanisms in the United States, maps them to established economic theory, and places them in comparative context for an international audience.

In the literature, there has been a tendency to ascribe differences in outcome solely to the use of a Mobile Party Pays regime (also know as a Receiving Party Pays regime). In this paper, we suggest that Mobile Party Pays is an important element, but that it needs to be understood in the context of other mechanisms that have had a complementary effect. Further, we argue that fixed and mobile termination rates need to be understood as a single integrated economic system.

---

[1] Author's current addresses: German Marshall Fund of the United States, Résidence Palace, Rue de la Loi 155 Wetstraat, 1040 Brussels, Belgium. E-mail: smarcus@GMFus.org. This author has affiliations with both the Federal Communications Commission (USA) and the European Commission, but the opinions expressed are solely those of the author, and do not necessarily reflect the views of either agency.

Contents

Abstract i

1. Introduction 1

2. An Overview of Call Termination in the United States 3
    2.1 Terminology and basic concepts 3
    2.2 Reciprocal Compensation 6
    2.3 The Terminating Monopoly 8
    2.4 Access charges and the terminating monopoly 9
    2.5 Reciprocal compensation rates between CLECs 12
    2.6 Termination rates of mobile operators 12
    2.7 The move to flat rate pricing 13
    2.8 Summary of reciprocal compensation and access charge arrangements 14
    2.9 The significance of symmetry 15

3. The U.S. mobile market in a global context 17
    3.1 Mobile penetration 18
    3.2 The cost of mobile services 19

4. Evolution of call termination in Europe and the U.S. 23
    4.1 Next steps for the European Union 23
    4.2 Next steps for the United States 24
    4.3 Concluding Remarks 25

Acknowledgments 27

Bibliography 28

Figures

Figure 1. The flow of reciprocal compensation. 4

Figure 2. The flow of access charges. 4

Figure 3. Canadian and U.S. Mobile Penetration Rates, 1990-2002. 19

Figure 4. Usage versus price per MoU for several developed countries. 20

Figure 5. ARPU for several developed countries. 22

Tables

Table 1. Reciprocal compensation. 14

Table 2. Access charges. 15

Table 3. Characteristics of mobile markets 18

1. Introduction

The economic framework under which the United States implements call termination fees[2] is unusual. Several recent studies suggest that the United States system has resulted in greater per-customer usage of mobile telephony services, and in lower average per-minute prices to consumers, than many other systems.[3] Our analysis supports these conclusions, but we do not believe that this is the end of the story.

Mobile termination rates have been a topic of intense debate in Europe in recent years. In the United States, termination rates are under challenge from a number of quarters. Technological and industry convergence, notably including IP telephony, is placing huge strains on the existing system. A reexamination of termination fee issues is timely.

Call termination fees tend not to be constrained by the competitive economic forces that constrain many other prices due to an effect (discussed later in this paper) known as the *terminating monopoly*. High and asymmetric call termination rates have raised concerns in Europe in recent years because they effectively force fixed users to provide large and arguably irrational subsidies to mobile users,[4] and also because they are one of several factors that contribute to European mobile average prices per *minute of use* (*MoU*) that are about twice as high as comparable prices in the United States.[5]

In the literature, there has been a tendency to ascribe these differences in usage and in price per MoU solely to the use in the U.S. of a Mobile Party Pays regime (also known as a Receiving Party Pays regime). We reject this notion as simplistic. In this paper, we suggest that Mobile Party Pays is an important element, but that it needs to be understood in the context of other mechanisms that have had a complementary effect.

---

[2] For purposes of this paper, we consider call termination fees to represent payments that one operator makes to another to complete a call, including not only reciprocal compensation but also access charges (defined in the next section). We do not distinguish between termination and *transit* (from the customer's central office to the point where the operators interconnect). In one instance (access fees), we discuss payments to the originating carrier, even though it is something of an oxymoron to refer to a payment to the originating operator as a call termination fee.

[3] Cf. Robert W. Crandall and J. Gregory Sidak, "Should Regulators Set Rates to Terminate Calls on Mobile Networks?" forthcoming in *Yale Journal on Regulation*, 2004. See also Stephen C. Littlechild, "Mobile Termination Charges: Calling Party Pays *vs* Receiving Party Pays", available at http://www.econ.cam.ac.uk/dae/repec/cam/pdf/cwpe0426.pdf. Finally, cf. FCC, *In the Matter of Implementation of Section 6002(b) of the Omnibus Budget Reconciliation Act of 1993, Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Service*, WT docket no. 02-379, released July 14, 2003 (hereinafter the *8th CMRS Competition Report*).

[4] The termination rates of the wired incumbent are typically limited by regulation to less than two eurocents per minute of use (European Commission, *Ninth report on the implementation of the EU electronic communications regulatory package* (hereinafter *9th Implementation Report*), COM(2003) 715 final, 19 November 2003), while European mobile termination rates average about 19 eurocents per minute of use (FCC, *8th CMRS Competition Report*). Thus, the subsidy flows from fixed operators to mobile operators.

[5] FCC, *8th CMRS Competition Report*. We return to this point later in this paper.

Further, we argue that fixed and mobile termination rates need to be understood as a single integrated economic system. Fixed-to-fixed, fixed-to-mobile, mobile-to-fixed, and mobile-to-mobile call termination rates interact in complex ways.

In particular, we argue that the U.S. intercarrier compensation regime for local calls, which establishes a presumption that local termination rates will be symmetric and based on the forward-looking costs of the incumbent wireline operator (unless the interconnecting party chooses to document a higher cost structure) has contributed strongly to low termination rates. This tendency to symmetry and low rates has permeated the system, even where regulation does not impose it. The termination rates for local calls between wireless operators, and also between non-incumbent wireline operators, are unregulated and are usually zero ("bill and keep").

The trend toward zero marginal *cost* for domestic U.S. calls has in turn fostered a migration to zero marginal retail *price*. Starting in 1998, wireless operators began offering nationwide "buckets of minutes" plans with no roaming or long distance charges. More recently, we are seeing the same evolution among wireline telephony operators.

Finally, we do not mean the paper merely to be a chest-thumping endorsement of United States regulatory policy, nor do we wish to naively suggest that other countries should rush to emulate our example. Nevertheless, the U.S. system does appear to have generated better results in a number of respects – perhaps as much through dumb luck as through regulatory genius. In any case, it is clear that the entire intercarrier compensation system will continue to face significant challenges in all countries in the years to come, and that further evolution is essential everywhere.

In support of that evolution, this paper seeks to summarize call termination fee mechanisms in the United States. We do not attempt to develop new economic theory; rather, we seek to map call termination fee mechanisms in the U.S. to established economic theory, and to place them in comparative context for an international audience, particularly a European audience.

This section provides the framework for the discussion that is to follow. The next section describes existing call termination fee mechanisms in the United States, and seeks to map the U.S. system to results in the economic literature. The subsequent section establishes a global context and compares the effects of call termination fee mechanisms in the U.S. to those of other developed countries in terms of mobile penetration and the cost of mobile service. We then offer concluding observations about the long-term challenges to the termination fee system, and prospects for future global evolution in Europe and the U.S.

2. An Overview of Call Termination in the United States

This section introduces key definitions and concepts, in keeping with our goal of making the system understandable to an international audience. It then proceeds with a description of call termination fee arrangements in the U.S., concluding with a tabular summary of the various mechanisms in place. Economic background is provided where appropriate, notably in regard to the termination monopoly problem. The section concludes with a discussion of the causes and implications of symmetry in call termination fees.

2.1 Terminology and basic concepts

Call termination arrangements in the United States depend on the nature of the call placed, and on the categorization of the carriers originating and terminating the call, in complicated ways. In the interest of simplifying the presentation we intentionally ignore some of the fine detail of the system;[6] unfortunately, it is impossible to fully grasp the system without mastering certain of its complexities.

Calls between two points in the same local calling area are *local calls*. Calls between two different areas are *long distance calls*. Carriers that provide local calling service over wired facilities are *local exchange carriers* (*LECs*). Carriers that provide long distance service are *interexchange carriers* (*IXCs*). Mobile operators provide *commercial mobile radio services* (*CMRS*).

Reciprocal compensation is associated with local calls; *access charges* are associated with long distance calls.

The boundaries of local calling areas (LATAs) do not correspond to those of states; thus, long distance calls may be either interstate or intrastate.

Historically, local telephone service was provided by monopoly operators; these local monopoly providers of wired telephone service are *incumbent local exchange carriers* (*ILECs*). In recent years, the market for local telephone service was opened to competition; the new entrants that compete with the ILECs in the provision of local calling service over wired facilities are *competitive local exchange carriers* (*CLECs*).

An FCC order further explains:

> "Existing intercarrier compensation rules may be categorized as follows: *access charge rules*, which govern the payments that interexchange carriers ("IXCs") and CMRS carriers make to LECs to originate and terminate long-distance calls; and *reciprocal compensation rules*, which govern the compensation between telecommunications carriers for the transport and termination of local traffic. Such an organization is clearly an oversimplification, however, as both sets of rules are subject to various exceptions …
>
> The access charge rules can be further broken down into *interstate* access charge rules that are set by this Commission, and *intrastate* access charge rules that are set by state

[6] For example, we ignore for the most part international calls, intrastate inter-LATA calls, and the Enhanced Service Provider exemption.

public utility commissions. Both the interstate and intrastate access charge rules establish charges that IXCs must pay to LECs when the LEC originates or terminates a call for an IXC, or transports a call to, or from, the IXC's point of presence ("POP"). CMRS carriers also pay access charges to LECs for CMRS-to-LEC traffic that is not considered local and hence not covered by the reciprocal compensation rules. … These access charges may have different rate structures—i.e., they may be flat-rated or traffic-sensitive. In general, where a long-distance call passes through a LEC circuit switch, a per-minute charge is assessed. …"[7]

Reciprocal compensation fees relate to local calls, and flow from the originating LEC to the terminating LEC (see Figure 1); access charges relate to long distance calls, and flow from the IXC to both the originating and terminating LECs (see Figure 2).

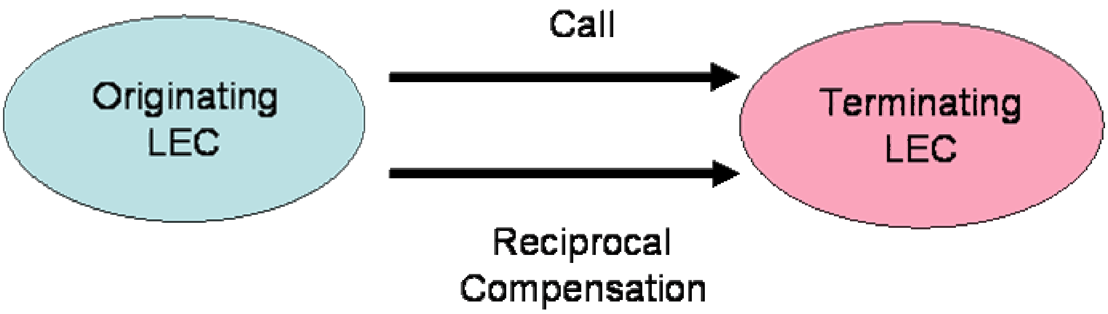

Call
Originating LEC
Terminating LEC
Reciprocal Compensation

Figure 1. The flow of reciprocal compensation.

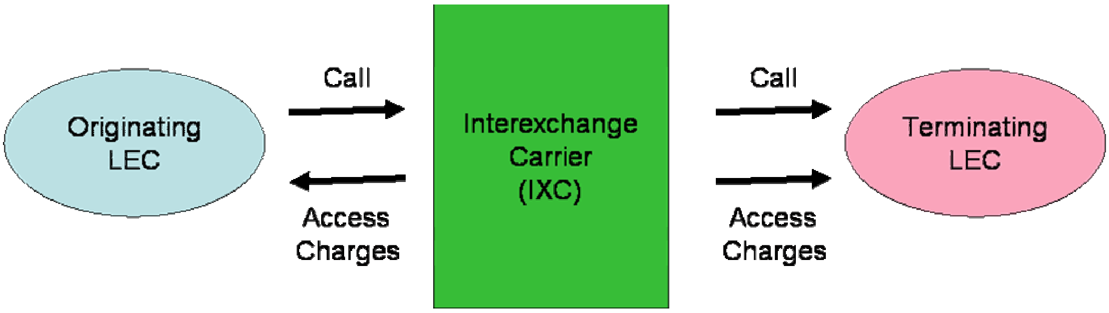

Originating LEC
Call
Access Charges
Interexchange Carrier (IXC)
Call
Access Charges
Terminating LEC

Figure 2. The flow of access charges.

In the discussion that follows, we will often refer to intercarrier compensation arrangements as *calling party's network pays* (*CPNP*), which reflects the widely implemented practice whereby the calling party's network pays a call termination fee to the network that terminates the call (and in the case of long distance calls, also to the LEC that originates the call). The literature often refers to these same arrangements as *calling party pays* (*CPP*), despite the fact that it is not really the calling party that pays. The fees of interest here flow between carriers, and do not necessarily correspond to retail payments by consumers.

---

[7] FCC, *In the Matter of developing a Unified Intercarrier Compensation Regime* (hereinafter *Unified Intercarrier Compensation NPRM*), CC Docket 01-92, released April 27, 2001, §§6-7.

The use of CPNP reflects the underlying assumption that the party that originates the call is the cost causer. This reflects in turn the underlying assumption that the originating party chooses to place the call, and is therefore willing to pay for the call, while the party that receives or terminates the call did not choose to receive it and is not necessarily willing to pay for the call. In recent years, many economists have called these assumptions into question. A recent paper by Jeon, Laffont and Tirole provides a detailed analysis.[8]

There is a tendency to speak of the U.S. as a *receiving party pays* (*RPP*) environment, or sometimes as *mobile party pays* (*MPP*) environment, in order to emphasize that it is not a CPNP environment. In fact, MPP and CPNP are not polar opposites. MPP refers to payments at the *retail* level. CPNP refers to intercarrier compensation in the form of call termination fees that flow from one carrier to another at the *wholesale* level.

The retail price of mobile services in the United States is unregulated. It is true that mobile operators generally account for minutes of use, whether for originating or for receiving calls,[9] but this is a commercial practice that is independent of call termination fees.

An extensive literature exists on call termination. Laffont, Rey and Tirole are generally credited with the definitive analysis.[10] A new paper by S.C. Littlechild[11] provides an extensive and thoughtful synthesis of the work on mobile termination to date.

---

[8] Doh-Shin Jeon, Jean-Jacques Laffont, and Jean Tirole, "On the receiver pays principle", to appear in the *RAND Journal of Economics*, 2004. They explore the inherent mirror-image relationship between calling and called party, and find that there is no qualitative difference, as "it takes two to tango." See also Crandall and Sidak, op. cit.; Littlechild, op. cit.. and FCC Working Paper 33, by Patrick DeGraba, "Bill and Keep at the Central Office As the Efficient Interconnection Regime", available at http://www.fcc.gov/osp/workingp.html.

[9] In many cases, mobile operators offer "buckets of minutes" plans where a consumer incurs no variable usage charges as long as usage is below an agreed quota of minutes; in this case, too, minutes of both origination and termination generally count against the quota, and are chargeable if they exceed the quota.

[10] See Jean-Jacques Laffont and Jean Tirole, *Competition in Telecommunications*, MIT Press, 2000. See also Jean-Jacques Laffont, Patrick Rey and Jean Tirole, "Network Competition: I. Overview and Nondiscriminatory Pricing" (1998a), *Rand Journal of Economics*, 29:1-37; and "Network Competition: II. Price Discrimination" (1998b), *Rand Journal of Economics*, 29:38-56. See also Crandall and Sidak, op. cit.; Littlechild, op. cit.; Michael Carter And Julian Wright, "Interconnection in Network Industries", *Review of Industrial Organization* 14: 1–25, 1999; Julian Wright, "Access Pricing under Competition: An Application to Cellular Networks", December 29, 2000; Doh-Shin Jeon, Jean-Jacques Laffont, and Jean Tirole, "On the receiver pays principle", to appear in the RAND Journal of Economics, 2004; Chris Doyle and Jennifer C. Smith, "Market Structure In Mobile Telecoms: Qualified Indirect Access And The Receiver Pays Principle", May 1999.

[11] Littlechild (2004), op. cit.

2.2 Reciprocal Compensation

Under the Communications Act[12], all LECs are required to establish reciprocal compensation arrangements for the transport and termination of telecommunications.[13] The Act establishes a preference that reciprocal compensation be addressed through voluntary negotiations between the carriers.[14]

In the event that the parties cannot agree, they may ask the relevant state commission to mediate any dispute, or (where at least one party is an ILEC) they may petition the state commission to arbitrate any open issues. In the context of an arbitration, the state commission is to consider the terms and conditions that an ILEC proposes[15] for such an agreement to be "just and reasonable" only to the extent that they result in the "mutual and reciprocal recovery by each carrier of costs associated with the transport and termination on each carrier's network facilities of calls that originate on the network facilities of the other carrier" based on a "reasonable approximation of the additional costs of terminating such calls."[16]

Carriers may choose to offset obligations in order to achieve "mutual recovery of costs", and are specifically permitted to waive mutual recovery altogether (e.g. to use "bill and keep" arrangements).[17] In other words, they can agree not to charge one another.

Under the FCC's implementing rules, when ILECs interconnect with non-dominant local carriers (be they wired or wireless) for the exchange of local traffic, the non-dominant carrier is presumed to have costs equivalent to those of the ILEC. This implies that reciprocal compensation rates will, by default, be symmetric. A non-dominant carrier retains the right to attempt to demonstrate underlying costs that are higher than those of the ILEC, but in practice this is rarely if ever done.

The combined effect of these provisions is that reciprocal compensation arrangements between an ILEC and any other wired or wireless carrier generally reflect either (1) the cost of the ILEC in both directions, or (2) no charges at all in either direction. In both cases, call termination fees are symmetric.

The FCC has summarized these arrangements in this way:

> Section 251(b)(5) [of the Communications Act of 1934 as amended, as codified at 47 U.S.C.] imposes on all [Local Exchange Carriers (LECs)] a "duty to establish reciprocal compensation arrangements for the transport and termination of telecommunications." Under current Commission rules interpreting the reciprocal compensation obligations of

---

[12] Communications Act of 1934 as amended, as codified at 47 U.S.C. (hereinafter *the Act*).

[13] 47 U.S.C. §251(b)(5).

[14] 47 U.S.C. §252.

[15] 47 U.S.C. §252(d)(2)(A).

[16] However, the rates are not necessarily the same as those TELRIC rates used to determine the price of Unbundled Network Elements (UNEs).

[17] 47 U.S.C. §252(d).

incumbent LECs [ILECs], the calling party's LEC must compensate the called party's LEC for the additional costs associated with transporting the call from the carriers' interconnection point to the called party's end office, and for the additional costs of terminating the call to the called party. The Commission's rules further require that the charges for both transport and termination must be set at forward-looking economic costs. The Commission's rules permit a state public utility commission ("PUC") to impose a bill-and-keep arrangement, provided that the traffic exchanged between the interconnecting carriers is relatively balanced and neither party has rebutted the presumption of symmetric rates.

> Existing access charge rules and the majority of existing reciprocal compensation agreements require the calling party's carrier, whether LEC, [interexchange carrier (IXC)] or [mobile], to compensate the called party's carrier for terminating the call. Hence, these interconnection regimes may be referred to as "calling-party's-network-pays" (or "CPNP"). Such CPNP arrangements, where the calling party's network pays to terminate a call, are clearly the dominant form of interconnection regulation in the United States and abroad. An alternative to such CPNP arrangements, however, is a "bill-and-keep" arrangement. Because there are no termination charges under a bill-and-keep arrangement, each carrier is required to recover the costs of termination (and origination) from its own end-user customers. As previously noted, under the Commission's rules, state PUCs may impose bill-and-keep arrangements on interconnection agreements involving an ILEC, provided that the traffic between the carriers is relatively balanced and neither carrier has rebutted the presumption of symmetrical rates. In addition, bill-and-keep arrangements are found in interconnection agreements between adjacent ILECs.[18]

It should be immediately apparent that:

- Reciprocal compensation termination fees between and ILEC and any other wireline or wireless carrier are on a Calling Party Network Pays (CPNP) basis. In this regard, they are not different from charges in most other countries. The rate of compensation is sometimes set to zero (bill-and-keep) by mutual agreement.
- Call termination fees are relevant at the wholesale level, but there is no regulatory requirement that they be flowed through to the retail level.
- ILECs are generally subject to retail rate regulation, at least for residential customers. The retail prices of other carriers (both mobile operators and wireline CLECs) are not regulated, so the degree to which retail prices reflect termination charges is a business decision, not a regulatory matter.
- For purposes of reciprocal compensation, mobile operators are generally treated no differently than competitive LECs (i.e. LECs that are not incumbent and thus presumed to be non-dominant).

---

[18] FCC, *In the Matter of Developing a Unified Intercarrier Compensation Regime* (hereinafter *Unified Intercarrier Compensation NPRM*), §§8-9, CC Docket 01-92, Notice of Proposed Rulemaking adopted April 19, 2001.

2.3 The Terminating Monopoly

> "So there I was, stranded with a broken down car in a one-horse town in Wisconsin with a gas station, a convenience store, and two barbers. I was on my way back from a two week fishing trip, and had to give an important talk in Chicago the next day. So what did I do? I did what any self-respecting, civilized man would do – I got a haircut!
>
> But here's the riddle. One barber looked pretty ragged, the other was well groomed. Naturally, I picked the one with the lousy haircut. And do you know why? I figured that he must have cut the hair of the barber who was well groomed! That was the man that I wanted – not the well-groomed barber, who presumably cut the hair of the barber who looked unkempt."[19]

The workings of call termination fee arrangements can be counterintuitive.

Call termination fees generally flow from the calling party's carrier to the receiving party's carrier.[20] As previously noted, the caller is presumed to be the cost causer.

This CPNP system tends to create perverse economic incentives. Carriers tend to be motivated to set termination rates vastly in excess of real costs, because in doing so they raise, not their own costs, but rather the costs of their rivals. To the extent that these costs are reflected in retail prices, they are reflected in the prices of their competitors, and not in their own prices.

Once a consumer subscribes to the carrier's service, that carrier controls a bottleneck that confers a degree of market power as regards calls that terminate to that customer. The market power arising from this bottleneck control is referred to as the *terminating monopoly*.

The market power arising from this bottleneck is exacerbated by the fact that, for a variety of practical and regulatory reasons, the consumer who places the call typically has at best limited visibility into the termination rates of the called party's operator. Regulation (for instance, geographic averaging requirements) may prevent the originating operator from flowing the full termination charge back to the consumer.[21] Users of pre-paid mobile service – which is rare in the U.S. but common elsewhere in the world – never see an itemized bill. The consumer may see only averaged call prices, or may not see individual call prices at all. For these reasons and others, the consumer who places the call typically lacks the economic signals that would enable him or her to seek to bypass high termination rates, and the consumer may have limited alternatives in any case.

The tendency toward above-cost termination rates is ultimately constrained by the price elasticity of demand. If a terminating operator increases its call termination rates, the increase may induce the firm's competitors to increase their retail prices. The increased prices will tend to depress retail demand for outgoing calls from the firm's competitors, including calls to the operator that initiated the process by increasing its rates. Unfortunately, the equilibrium price in such a system is likely to be much higher than the actual call termination cost to the carrier that sets the high termination rate, and the equilibrium demand for calls to that carrier correspondingly lower than that which would exist absent the terminating monopoly.

---

[19] Derived from an old joke.

[20] By definition, call termination fees are for termination. Note that access charges can also flow to the *originating* LEC.

[21] Cf. DeGraba, op. cit., page 8: "… carriers with smaller market shares may have a greater incentive to charge excessive terminating access charges because those charges are unlikely to be flowed through to interconnecting carriers' end-user prices."

Returning to our metaphor, we can now explain why the shaggy barber should be preferred. High termination rates do not directly raise costs to the customers of the operator that sets them; rather, *they tend to raise costs to those who place calls to that operator's customers*. They impact the prices of an operator's competitors, not those of the operator itself.

A recent paper by Haucap and Dewenter[22] is particularly relevant. They study call termination rates in a CPNP system where the calling party has little or no visibility into termination fees (as is often the case for the reasons previously noted). They develop a mathematical model that provides two key insights into termination. First, they find that operators with a small number of customers tend to set termination rates *higher* than those with a large number of customers (because the rates that small carriers set have less impact on the average price paid by their competitors' customers). Second, they find that a regulatory "cap" solely on the termination rate of operators with market power in their respective home markets may serve to exacerbate, rather than to ameliorate, the problem of termination rates that greatly exceed costs.

Haucap and Dewenter use a regression analysis, based on termination rates of European carriers, to validate their model. They find a statistically significant negative correlation (in other words, a correlation in the predicted direction) between termination rate and number of subscribers. Interestingly, they find no significant correlation between termination rate and the HHI associated with the operator's home market. These findings are consistent with the notion an operator need not have Significant Market Power (SMP) in a retail market in order to be motivated to impose elevated termination costs; indeed, operators with high market shares will tend to be more constrained by the prospect of reducing the total call volume (due to demand elasticity to the extent that high call termination rates are reflected in retail prices).

2.4 Access charges and the terminating monopoly

With that theoretical background out of the way, we now return to call termination in the United States. High call termination rates have raised concerns in recent years in the mobile environment; however, the relevant economic models are not specific to the mobile market.

Recall that an IXC pays access charges to both the originating and the terminating LEC. Where the terminating LEC is a large ILEC, access charges are set in the range of

---

[22] Dewenter, Ralf, and Haucap, Justus, "Mobile Termination with Asymmetric Networks", October 2003, available via SSRN. Presented at the 4th ZEW Conference on the Economics of Information and Communication Technologies, Mannheim, Germany, July 2004.

$0.0055 to $0.0065.[23] Slightly higher rates are permitted for certain smaller ILECs, including certain rural rate-of-return operators.[24]

When the Telecommunications Act of 1996 opened local markets to competition, the FCC did not initially regulate the access charges that CLECs would assess on IXCs to originate and terminate calls. At the same time, because of statutory rate averaging requirements, IXCs were prohibited from charging different retail rates, even if access charges differed. As a result, numerous competitive local exchange carriers (CLECs) began to charge extremely high originating and terminating charges. In other words, these regulatory provisions had a net effect analogous to that studied by Haucap and Dewenter: they established a Calling Party's Network Pays system, they reinforced the terminating monopoly power of the CLEC, and they blocked customer visibility into the relevant pricing signals that might have enabled customers to respond.

In an order issued in 2001, the FCC summarized the problem as follows:

> Despite previous indications that market forces might constrain CLEC access rates, the Commission recently found that, in actuality, the market for access services is not structured in a manner that allows competition to discipline rates. Specifically, the Commission found that the originating and terminating access markets consist of a series of bottleneck monopolies over access to each individual end user. Once an end user decides to take service from a particular LEC, that LEC controls an essential component of the wireline system that provides interexchange calls, and it becomes a bottleneck for IXCs wishing to complete calls to, or carry calls from, that end user. Thus, with respect to access to their own end users, CLECs have just as much market power as ILECs. In addition, the Commission determined that "the combination of the market's failure to constrain CLEC access rates, the Commission's geographic rate averaging rules for IXCs, the absence of effective limits on CLEC rates and the tariff system created an arbitrage opportunity for CLECs to charge unreasonable access rates." … Because the CLEC access market is not truly competitive, we cannot simply assume that ``whatever the market will bear'' translates into a just and reasonable rate.[25]

The magnitude of the disparity in termination costs was quite significant:

> The access rates charged by ILECs operating in BTI's service areas are a relevant benchmark, because ILEC switched access services are functionally equivalent to CLEC switched access services. In addition, according to fundamental economic principles, in a properly functioning competitive market, the access rates of BTI's primary access competitors would have been a substantial factor in BTI's setting of its own access rates. Indeed, in other markets, BTI's pricing behavior adhered to these principles. BTI's rates for its local exchange service were approximately 15 to 25 percent below those of its primary competitors, BellSouth and GTE; and BTI's rates for long distance service were roughly the same as those of its primary IXC competitors.

---

[23] These provisions specifically apply to a class of large ILECs regulated pursuant to price caps. Smaller ILECs are generally regulated on a rate-of-return basis.

[24] FCC, "In the Matter of Access Charge Reform, Price Cap Performance Review for Local Exchange Carriers, Low-Volume Long-Distance Users, Federal-State Joint Board On Universal Service" ("CALLS Order"), CC Dockets 96-262, 94-1, 99-249, 96-45, released May 31, 2000.

[25] FCC, *In the Matters AT&T Corp., Complainant, versus Business Telecom, Inc., Defendant. Sprint Communications Company, L.P., Complainant, Business Telecom, Inc., Defendant,* Section III.B.1. Memorandum Opinion And Order, EB Docket EB-01-MD-001, Released: May 30, 2001.

> Nevertheless, during all relevant times, BTI's access rate was significantly higher than the competing ILECs' rates. In July 2000, BTI's access rate of 7.1823 cents per minute was more than 15 times higher than BellSouth's average rate of approximately 0.48 cents per minute, and more than 7 times higher than GTE's average rate of approximately 1.0 cent per minute. In July 1999, BTI's access rate was more than 5 times higher than BellSouth's average rate of approximately 1.4 cents per minute, and more than 3.5 times higher than GTE's average rate of approximately 2.0 cents per minute. In July 1998, BTI's access rate was approximately 4.5 times higher than BellSouth's average rate of approximately 1.6 cents per minute, and more than 2.5 times higher than GTE's average rate of approximately 2.8 cents per minute. [26]

The access charges that BTI, a CLEC, imposed on AT&T as recently as 2000 were thus in excess of 7 cents per minute, while charges that BellSouth imposed on AT&T in the same areas at the same were about a half cent per minute. The disparity is striking. The ratio is comparable to that between European mobile termination charges (about $0.19 per minute of use)[27] versus European fixed termination rates (a bit less than $0.02 per minute).[28] In both cases, the ratio is between a service with a termination monopoly and no regulatory constraint, on the one hand, and a regulated wireline incumbent operator on the other.

The cases are not strictly comparable – access charges are somewhat different from reciprocal compensation charges between local carriers. Recall that reciprocal compensation charges (see Figure 1) flow in either direction: When carrier A is the originating carrier, carrier A pays terminating local carrier B; when however A terminates a call originated by B, then B pays A. Access charges, however, are one-way charges – it is always the IXC that pays. The IXC pays both the originating local carrier and the terminating local carrier (see Figure 2).

The FCC found it necessary to regulate CLEC access charges by imposing a "cap", based on the regulated access charges of the adjacent ILEC.[29] CLECs may unilaterally establish access charge rates by tariff as long as they are below the cap. If they wish to establish access charge rates above the cap, they must do so through voluntary negotiations.

There are some striking parallels between the reciprocal compensation rules and the access charge rules. First, it is the incumbent LEC that establishes the presumptive default rate for both reciprocal compensation and for access charges. In both cases, CLEC charges are not individually regulated, but the rates will tend in general to be the same as those of the corresponding ILEC. Finally, the CLEC is not required to provide cost data to regulators.

---

[26] Ibid., section III.B.2.a.

[27] FCC, *8th CMRS competition report*, 2003.

[28] European Commission, *9th implementation report*.

[29] FCC, *In the Matter of Access Charge Reform*, CC Docket No. 96-262, FCC No. 01-146 (rel. Apr. 27, 2001), at 45. "Our orders addressing ILEC access charges have consistently stated our preference to rely on market forces as a means of reducing access charges. Thus, in setting the level of our benchmark, we seek, to the extent possible, to mimic the actions of a competitive marketplace, in which new entrants typically price their product at or below the level of the incumbent provider. We conclude that the benchmark rate, above which a CLEC may not tariff, should eventually be equivalent to the switched access rate of the incumbent provider operating in the CLEC's service area… We also adopt rules to ensure that no CLEC avails itself of our benchmark scheme to increase its access rates, and we adopt a separate benchmark for certain firms operating in rural areas."

2.5 Reciprocal compensation rates between CLECs

Interconnection of CLECs with ILECs (reciprocal compensation) and with IXCs (access charges) has been previously addressed.

When CLECs interconnect with one another, the rate of reciprocal compensation is unregulated. It is a matter of private negotiation. CLECs can choose to adopt a bill-and-keep regime, which is to say that they can set a reciprocal compensation rate to one another of zero.[30]

2.6 Termination rates of mobile operators

In implementing the reciprocal compensation provisions of the 1996 Act, the FCC treated mobile operators as if they were CLECs. Thus, when a mobile operator interconnects locally with an ILEC, reciprocal compensation flows from the originating carrier to the terminating carrier. Moreover, there is a presumption that the call termination rates will be symmetric based on the forward looking costs of the ILEC.[31] Mobile operators, like other CLECs, have the option of demonstrating that their higher traffic-sensitive termination costs entitle them to a higher, asymmetric termination rate.

When a mobile operator interconnects locally with another mobile operator, or locally with a CLEC, the rate for reciprocal compensation is established through unregulated commercial negotiation. These agreements are generally on a bill-and-keep basis.[32]

When a mobile operator originates a long distance call, it generally establishes a contractual resale relationship with a long distance carrier. Access fees are not relevant.[33]

Mobile operators are not permitted to establish tariffs for access charges where they terminate long distance calls from IXCs. They could, in principle, voluntarily negotiate a compensation rate with an IXC, but this rarely happens. In practice, where a mobile operator terminates a long distance call, it is generally on a bill-and-keep basis (no money changes hands).[34]

---

[30] FCC, *Unified Intercarrier Compensation NPRM*.

[31] For a lengthy discussion of the nuances of mobile-LEC interconnection, see the *Unified Intercarrier Compensation NPRM*, §§78-95. In essence, mobile-LEC interconnection is regulated under §§251-252 of the Communications Act, just as is LEC-LEC interconnection.

[32] Ibid., §95.

[33] Ibid., §96. Now that Regional Bell Operating Companies (RBOCs) are permitted to offer long distance service, they usually adopt similar arrangements for origination of long distance calls.

2.7 The move to flat rate pricing

The typical European pattern is one of Calling Party's Network Pays, with mobile termination rates that averaged about of $0.19 per minute of use.[35] Not surprisingly, the retail price in these countries generally exceeds the termination rate, which the carrier views as a cost. These high per minute costs tend to make it difficult for carriers to offer flat rate calling plans.[36] A flat rate plan would have to address many business risks, including the prospect that the plan might attract large numbers of self-selected customers who had significantly above-average usage patterns.

Conversely, call termination rates in the United States that are less than $0.01 in most cases, and zero in many cases, facilitate flat rate pricing.

AT&T Wireless's offer of Digital One Rate in 1998 represents a watershed event in this regard. AT&T offered a plan with flat rates across the United States. As long as the customer used not more than some fixed (and possibly large) number of minutes of air time, the customer could place or receive calls to and from any point in the continental United States. The customer would incur no per-minute charges, no long distance charges, and no roaming charges.[37]

Not surprisingly, Digital One Rate was immensely popular. The success of Digital One Rate effectively forced its mobile competitors to provide a competitive response; however, initially they were hampered by their lack of nationwide scale. The net result was a wave of consolidation, alliances and joint ventures that ultimately resulted in a nationwide market for mobile telephone services with multiple carriers, each offering nationwide plans offering a large bucket of minutes for a flat monthly fee.

---

[34] Cf. Ibid., §94.

[35] European Commission, *9th Implementation Report*, page 18. The figure is for SMP operators, effective August 2003. Euro prices are converted to dollars (here and throughout this paper) at an assumed exchange rate of $1.20 per Euro. Cf. FCC, *8th CMRS competition report*, at 207.

[36] Cf. Laffont and Tirole, *Competition in Telecommunications*, page 190: "It is correct that a change in the access charge need not affect the (absence of) net payment between the operators, but the access charge affects each network's perceived marginal cost and therefore retail prices." See also DeGraba, op. cit., page 8: "… because carriers will view traffic-sensitive interconnection charges as raising their marginal costs, they will tend to raise their traffic-sensitive retail prices, even though the underlying cost structure of the networks may be non-traffic-sensitive."

[37] Cf. *8th CMRS Competition Report*, §94: "AT&T Wireless's Digital One Rate ("DOR") plan, introduced in May 1998, is one notable example of an independent pricing action that altered the market and benefited consumers. Today all of the nationwide operators offer some version of DOR pricing plan which customers can purchase a bucket of MOUs to use on a nationwide or nearly nationwide network without incurring roaming or long distance charges." Several mobile operators offer a variant of this plan where there are no roaming charges as long as the customer is using that operator's facilities.

One dramatic result has been a reduction in roaming charges. While roaming charges comprised 14% of mobile revenues in 1995[38], they represented just 5% of mobile revenues in 2002, and 4% in 2003[39].

Today, flat rate plans are becoming increasingly prevalent for all forms of telephony.[40] As ILECs were permitted to offer long distance services, they typically offered flat rate plans with unlimited domestic long distance.[41] Traditional long distance carriers offer combined local and long distance service at a flat rate.[42] IP telephony service providers commonly offer unlimited domestic calls at a flat rate.[43]

2.8 Summary of reciprocal compensation and access charge arrangements

Reciprocal compensation arrangements are graphically summarized in Table 1, where codes A and B are explained below:

Origination	Termination		
	ILEC	CLEC	Mobile
ILEC	A	B	B
CLEC	B	A	A
Mobile	B	A	A

Table 1. Reciprocal compensation.

A – Terms are established through voluntary negotiations, often as bill-and-keep.

---

[38] Cellular Telecommunications and Internet Association, *Semi-Annual Wireless Industry Survey* (see http://www.wow-com.com/industry/stats/surveys/).

[39] Ibid.

[40] These flat rate plans are truly flat rate, whereas the mobile plans are generally two part tariffs. The usage charges of the mobile plans are usually set to very high levels (in the range of $0.40 per MoU). They are not so much intended to be used, as to punish consumers who purchase bundles that are too small. The common feature between the mobile plans and the newer truly flat rate plans is a movement away from meaningful usage charges.

[41] Verizon, for example, offers 1,000 minutes of long distance service for prices in the range of forty dollars per month. See: http://www22.verizon.com/ForYourHome/sas/res_fam_LongDistancePlans.asp.

[42] See, for instance, http://www.consumer.att.com/plans/bundles/. Prices in the range of $49.95 for local service plus unlimited domestic long distance are not uncommon.

[43] For example, Vonage offers unlimited calls to or from the U.S. and Canada for $29.99 a month. See www.vonage.com.

B – Where the terminating operator is an ILEC, reciprocal compensation is paid to the ILEC at a rate based on the ILEC's forward looking marginal cost. Where the terminating operator is a CLEC or mobile operator, reciprocal compensation is paid to the CLEC or mobile operator at a rate based on the ILEC's forward looking marginal cost unless the CLEC or mobile operator can demonstrate a higher forward looking marginal cost.

Access charge arrangements flow from the IXC to the operator associated with origination or termination of the call. These arrangements are summarized in Table 2 below. The left column represents access charges due to the originating operator, while the right column represents access charges due to the terminating operator.

	Origination	Termination
ILEC	C	C
CLEC	D	D
Mobile	E	F

Table 2. Access charges.

C – Access charges are due to an originating or a terminating ILEC in accordance with CALLS, at rates limited to \$0.0055-\$0.0065 for (large) rate cap LECs. Somewhat higher rates are permitted for (small or rural) rate-of-return LECs.

D – Access charges may be tariffed by an originating or a terminating CLEC at rates up to those of the corresponding ILEC, unless a higher rate is voluntarily agreed.

E – The originating mobile operator usually contracts with the IXC to resell minutes, so access charges are irrelevant.

F – No access charges are payable to a terminating mobile operator unless the parties agree otherwise.

2.9 The significance of symmetry

As we have seen, the call termination system in the U.S. has a strong tendency toward symmetry in the rates charged for reciprocal compensation, and toward identical access charge rates for wired carriers in the same geographic area (whether ILEC or CLEC). These characteristics serve to prevent many forms of regulatory arbitrage, including exploitation of the terminating monopoly.

ILEC-CLEC and ILEC-CMRS reciprocal compensation rates are generally symmetric, and set at a rate that reflects the marginal cost of the ILEC.

ILEC-ILEC, CLEC-CLEC, CLEC-CMRS, and CMRS-CMRS reciprocal compensation rates are determined through voluntary negotiations, and in many cases are set to zero (bill-and-keep). ILEC-ILEC and CMRS-CMRS interconnection is usually on a bill-and-keep basis.[44] Traffic patterns are usually in rough balance in these cases; consequently, not much money is likely to change hands between the carriers due to reciprocal compensation.[45] The carriers presumably choose to minimize transaction costs by avoiding the need to account for traffic and deal with disputes. The zero rate also avoids the business risk associated with the possibility that the balance of traffic might shift in an unfavorable direction over time. The mitigation of this risk serves in turn to facilitate the use of flat rate pricing.

If traffic were significantly imbalanced, voluntary negotiations of symmetric rates would not lead to low or zero rates. The carrier that terminates more calls than it originates would prefer a high rate, while the carrier that originates more call than it terminates would prefer a low or zero rate. Carriers in developing countries tend to terminate far more calls from carriers in developed countries than they originate. Under these circumstances, carriers in developing countries will *ceteris paribus* tend to prefer high call termination rates (which can be orders of magnitude in excess of marginal cost) over low or zero rates.

The presumptions of symmetry in reciprocal compensation rates, and of CLEC-ILEC parity in access charge rates, also serve to reduce regulatory burdens. ILECs must offer call termination to CLECs and mobile operators at rates based on the ILEC's forward looking costs.[46] CLECs and mobile operators need not cost-justify their rates, since their rates are routinely based on those of the ILEC.[47]

The presumption of symmetry has important consequences. In most European countries, large asymmetries in termination rates exist between wired carriers (who are typically subject to termination rate regulation) and mobile operators (who historically have not been subject to termination rate regulation). Rates often differ by an order of magnitude.[48] This asymmetry has effectively transferred billions of dollars from fixed operators to mobile, creating an irrational subsidy. The U.S. has avoided this market distortion, largely through the use of symmetric call termination rates.

---

[44] FCC, *Unified Intercarrier Compensation NPRM*, at 9 and 95.

[45] Note, however, that the level of charging will tend to affect their perception of marginal cost, and is thus likely to influence their pricing decisions. See Laffont, Rey and Tirole (1998a), and also Laffont and Tirole, *Competition in Telecommunications*, page 190.

[46] The access charge rates established by CALLS are claimed to correspond approximately to cost-based rates.

[47] CLECs and mobile operators have the prerogative to attempt to justify a higher reciprocal compensation rate based on costs higher than those of the ILEC, but this is rarely done in practice.

[48] European Commission, *9th Implementation Report*, page 18: "… although there has been a decrease in interconnection charges, their level remains on average more than 9 times higher than fixed-to-fixed interconnection charges (double transit)."

3. The U.S. mobile market in a global context

This section of the paper evaluates the effectiveness of the U.S. call termination system in terms of its impact on the marketplace. We confine ourselves to the mobile marketplace because it is in regard to mobile telephony that the U.S. call termination system is conspicuously different from that of other countries, and also because the marketplace differences between the U.S. and other countries are more dramatic for mobile telephony than for fixed.

The low or zero termination fees that exist in the United States tend to facilitate flat rate mobile pricing. By contrast, high mobile termination rates in Europe and elsewhere tend to enforce high charges per mobile minute of use, but also support low initial cost for mobile service (due to handset subsidies, pre-paid calling card plans, and other incentives to consumers).

The relative impact is as might be anticipated: the European pattern has encouraged rapid adoption of mobile telephone service, but has also had a tendency to depress usage of those phones (expressed in minutes of use per month). Conversely, the U.S. approach has led to slower adoption of mobile telephone service, but has encouraged much higher utilization of mobile phones.[49]

[49] Cf. Crandall and Sidak: "Mobile subscribers in MPP countries appear to use their mobile phones more intensively, presumably because of the pricing structure that MPP elicits from competitive MNOs."

3.1 Mobile penetration

Table 3[50] is a widely cited comparison of mobile penetration, usage, and revenue per minute in several leading global economies, as of late 2002.

Country	CPP or MPP	Penetration (%)	Share of Prepaid (%)	MOUs	Revenue per Minute ($)
USA	**MPP**	**49**	**5**	**458**	**0.12**
Canada	**MPP**	**37**		**270**	**0.11**
UK	**CPP**	**85**	**69**	**132**	**0.22**
Germany	**CPP**	**72**	**54**	**72**	**0.29**
Italy	**CPP**	**93**		**121**	**0.20**
France	**CPP**	**63**		**156**	**0.20**
Finland	**CPP**	**85**		**146**	**0.24**
Japan	**CPP**	**62**	**3**	**170**	**0.30**
South Korea	**CPP**	**68**	**1**	**296**	**0.10**
Australia	**CPP**	**68**		**173**	**0.16**

Table 3. Characteristics of mobile markets.

Crandall and Sidak have analyzed the underlying penetration data, particularly as regards Canada and the United States (see Figure 3), and have found that "…if the growth rate continues to follow this S-shape pattern, mobile penetration in the United States should equal the penetration rates realized in most CPP countries between 2008 and 2014. The growth in mobile subscribers in Canada is similarly impressive—26.8 percent in 2000, 22.3 percent in 2001, and 11.8 percent growth in 2002. … [M]obile penetration in Canada and the United States will likely equal the penetration rates of CPP countries in the near term …"[51]

---

[50] FCC, *8th CMRS Competition Report*, Table 12. Cited data sources are: Linda Mutschler, *Global Wireless Matrix 4Q02*, Global Securities Research, Merrill Lynch, Apr. 2, 2003; and Linda Mutschler, *The Next Generation VII*, Global Securities Research, Merrill Lynch, Feb. 21, 2003. Per the *8th CMRS Competition Report*, "average MOUs include both incoming and outgoing traffic, and usually exclude traffic related to mobile data services".

[51] Sidak, J. Gregory and Crandall, Robert, "Should Regulators Set Rates to Terminate Calls on Mobile Networks?" *Yale Journal on Regulation*, Vol. 21, 2004, http://ssrn.com/abstract=462041.

.

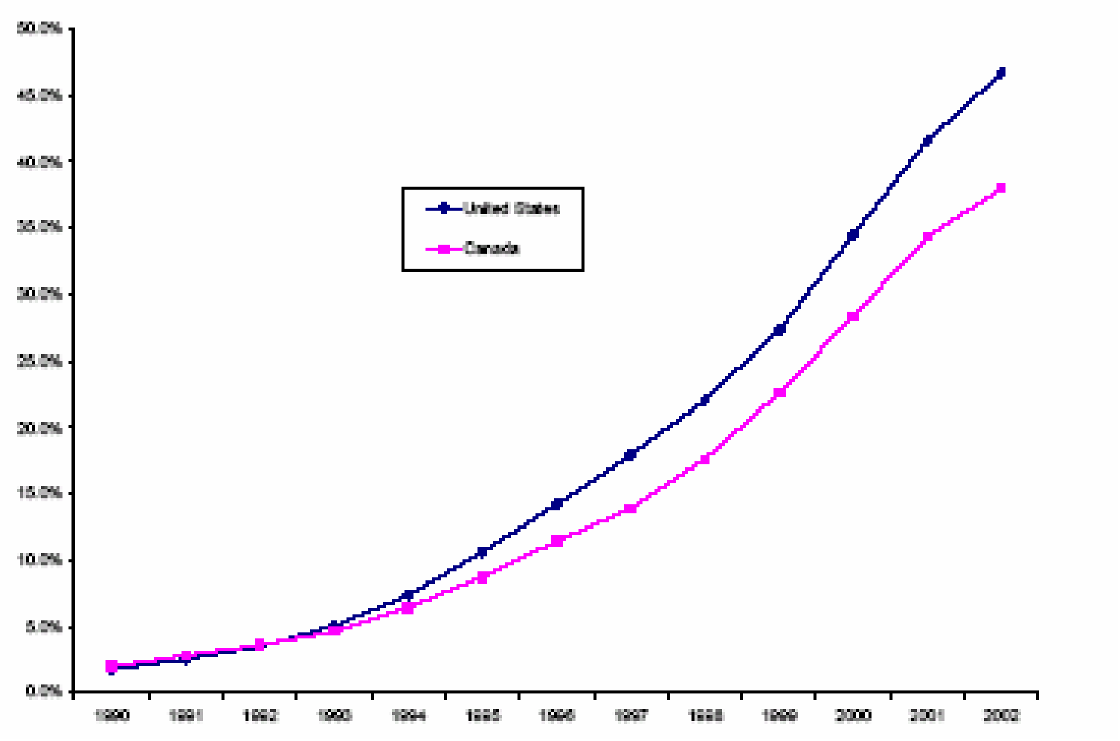

United States
Canada
50.0%
45.0%
40.0%
35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002

Figure 3. Canadian and U.S. Mobile Penetration Rates, 1990-2002.[52]

3.2 The cost of mobile services

Figure 4 is a restatement of Table 3. It relates the cost per minute (in U.S. dollars) of mobile usage to average Minutes of Use (MoU) per month, based on the data in Table 3.[53] In fitting a regression curve to the data, we have somewhat arbitrarily assumed a linear relationship. The data show the expected negative correlation.

---

[52] Crandall and Sidak, op. cit.

[53] Again, per the *8th CMRS Competition Report*, "average MOUs include both incoming and outgoing traffic, and usually exclude traffic related to mobile data services".

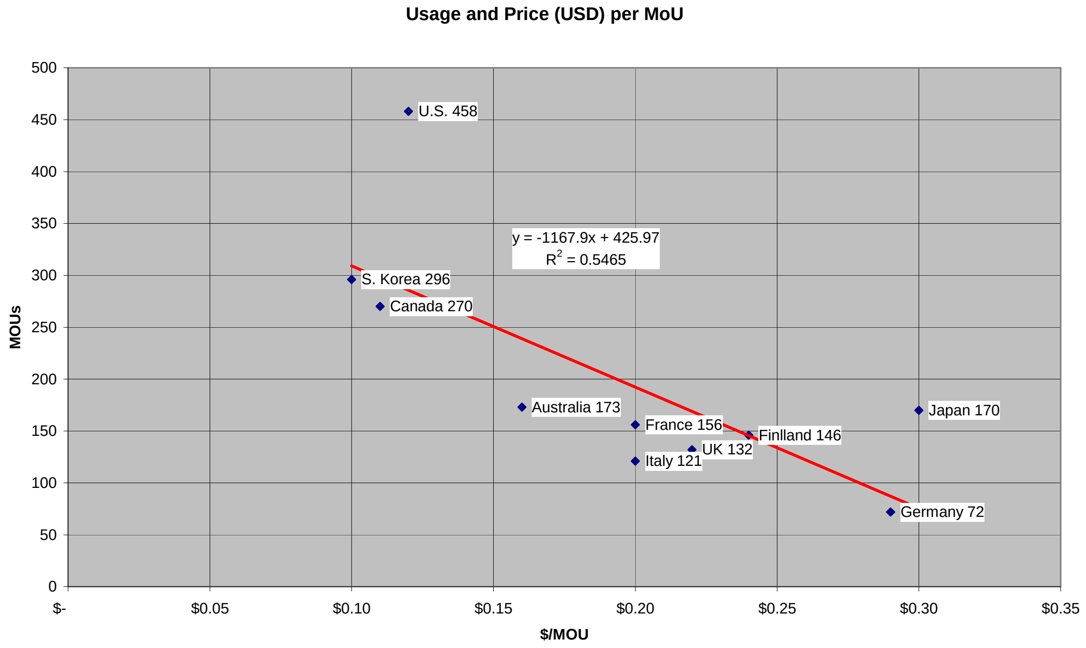

Usage and Price (USD) per MoU
MOUs
$/MOU
500
450
400
350
300
250
200
150
100
50
0
$-
$0.05
$0.10
$0.15
$0.20
$0.25
$0.30
$0.35
U.S. 458
y = -1167.9x + 425.97
R2 = 0.5465
S. Korea 296
Canada 270
Australia 173
France 156
Finlland 146
UK 132
Italy 121
Japan 170
Germany 72

Figure 4. Usage versus price per MoU for several developed countries.

Tempting as it might be to interpret the downward slope as corresponding to demand elasticity, it is not formally correct to do so. These are not the same consumers. A consumer in France cannot in reality switch to a Korean mobile phone operator. Furthermore, there are significant differences among these countries as regards calling preferences, and also as regards disposable income. Nonetheless, it is fair to say that the data suggest that demand is elastic.

It is often instructive to examine the outliers and residuals of a regression curve. In this case, the United States and Germany represent interesting cases.

At 458 MoU, the United States demonstrates considerably more usage than its average price per minute might otherwise lead us to expect. This probably reflects consumer response to buckets-of-minutes plans: "[A] a U.S. mobile subscriber who opts for a large bundle of minutes with virtually unlimited night and weekend minutes perceives that the incremental price of using a wireless minute is virtually free, whereas a mobile subscriber in the U.K. does not have the same perception."[54]

It is perhaps noteworthy that the U.S. experiences much higher MoU than either South Korea or Canada, even though the average prices per MoU of all three are similar. Equally intriguing is the similarity in price per minute and the MoUs between South Korea and Canada, even though the former is a CPP environment, and the latter an MPP environment. To the extent that the MoUs represent consumer response to perceived pricing, this is not so surprising. We could reasonably expect that consumers would

[54] FCC 8th CMRS Competition Report, §204.

respond to the price per minute, which they experience directly, and not to the CPP/MPP distinction, which is not directly visible to them.

Germany poses more of a riddle. Price per MoU is much higher than that of its European neighbors, so it is not surprising that MoUs consumed per month are much lower than those in many European countries. When one considers the Average Revenue per User (ARPU), the difference becomes particularly striking (see Figure 5). For any point in Figure 4, the associated ARPU is simply the area under the rectangle that the point forms with the origin (i.e. the product of MoUs and price per MoU). Germany's rectangle is long and low, and its ARPU is consequently significantly less than that of many of its European neighbors.

Termination rates alone cannot explain this anomaly. German termination rates are in the range of $0.17 per minute,[55] not very different from the European average of about $0.19 per minute.[56] They cannot fully explain retail prices in the neighborhood of $0.29 per minute.

Analysis by Sanford Bernstein supports these findings.[57] "Germany continues to leave value on the table relative to the other 4 Major European markets in service revenue per pop and end customer spending per pop… Based on 2003 levels, Germany's monthly service revenue per pop is €18.7 compared with a €25.7 average for the rest of Europe (27% smaller). Germany's monthly end customer spending per pop is also smaller (23%) than the average of the other 4 Major European markets (€15.8 versus €20.4)."

In a competitive market, and assuming that demand is reasonably elastic, one might normally expect that German mobile operators would find it profitable to lower the price per MoU in order to increase ARPU to levels more comparable to those of other European countries. The data suggest that this might generate a very substantial increase in revenue. Why does this not happen? Is the market less competitive than might be expected, or are other, more subtle factors at work?[58] Or is this simply a case, as one market participant has suggested, where the players know what they need to do, but have not yet found the right way to implement and market their services?

---

[55] Arno Wirzenius (for the Ministry of Transport and Communications, Helsinki), Mobile Pricing and Interconnection Regimes, 17 May 2004. See page 12. Price is net of VAT. Euro prices are converted to dollars (here and throughout this paper) at an assumed exchange rate of $1.20 per Euro.

[56] European Commission, *9th Implementation Report*, page 18. The figure is for SMP operators, effective August 2003. Euro prices are converted to dollars (here and throughout this paper) at an assumed exchange rate of $1.20 per Euro. Cf. FCC, *8th CMRS competition report*, at 207.

[57] Andrew J. O'Neill and J. Kyle Raver, "Euro Wireless: New T-Mobile contract bundles to start catch-up of German wireless spend to average European levels", 16 January 2004.

[58] The same Bernstein analysis observes that the two large operators in this market have spectrum limitations, and may therefore perceive high costs. This cannot fully explain the anomaly, but it may be a contributing factor.

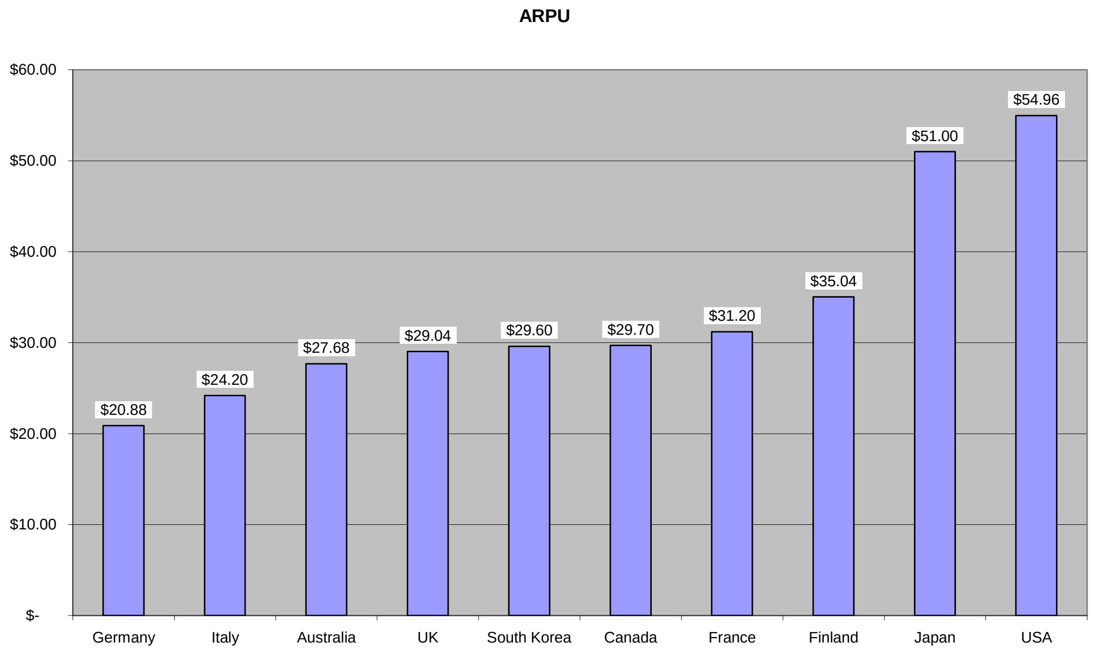

ARPU
$60.00
$50.00
$40.00
$30.00
$20.00
$10.00
$-
$20.88
$24.20
$27.68
$29.04
$29.60
$29.70
$31.20
$35.04
$51.00
$54.96
Germany
Italy
Australia
UK
South Korea
Canada
France
Finland
Japan
USA

Figure 5. ARPU for several developed countries.

4. Evolution of call termination in Europe and the U.S.

This section of the paper briefly contrasts current and projected developments in the European Union to those in the United States. It closes with some brief comparative observations.

4.1 Next steps for the European Union

The European Union is in the process of implementing a New Regulatory Framework (NRF) for electronic communications.[59] Under the NRF, the European Commission identifies a number of markets where carriers are likely to possess SMP and where, accordingly, *ex ante* regulation (i.e. regulation prior to a competitive abuse) may be appropriate. National Regulatory Authorities (NRAs) interpret those markets in terms of their national circumstances, identify firms (if any) that have Significant Market Power (SMP) on those markets, and apply minimal "proportionate" remedies to address the harms that SMP is likely to engender.

The Commission has addressed the call termination problem through the market definition mechanism. The Commission has identified eighteen markets that are potentially amenable to *ex ante* regulation. Among these are markets for call termination *to the customers of an individual fixed or mobile operator*.[60] Defining the market in this way will tend to create a strong presumption of SMP in regard to termination of calls for that operator's own customers, unless rebutted by specific facts. If SMP is found, the NRA determines what regulatory remedies are appropriate. This process may eventually lead to cost-based termination rates for far more carriers than are presently subject to them.

This overall approach is logical, and is in fact the most natural way to deal with high termination fees under the NRF. At the same time, it will tend to lead to highly regulated outcomes. In the cases that have been notified[61], it has generally led to heavy regulatory controls, including cost-orientation for termination rates.

---

[59] See J. Scott Marcus, Federal Communications Commission (FCC) Office of Strategic Planning and Policy Analysis (OSP) Working Paper 36, "The Potential Relevance to the United States of the European Union's Newly Adopted Regulatory Framework for Telecommunications," July 2002, available at http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-224213A2.pdf. The article and derivative works also appear in: *Rethinking Rights and Regulations: Institutional Responses to New Communications Technologies*, Ed. Lorrie Faith Cranor and Steven S. Wildman, MIT Press, 2003; in the *Journal on Telecommunications and High Technology Law* 111 (2003); and in the *2004 Annual Review* of the European Competitive Telecommunications Association (ECTA). The relevant European Directives appear in the bibliography of this paper.

[60] *Commission Recommendation of 11 February 2003 on relevant product and service markets within the electronic communications sector susceptible to ex ante regulation in accordance with Directive 2002/21/EC of the European Parliament and of the Council on a common regulatory framework for electronic communications networks and services*, Official Journal of the European Communities, L 114, (2003/311/EC), May 8, 2003. Market 9 is "Call termination on individual public telephone networks provided at a fixed location"; market 16 is "Call termination on individual mobile networks."

For now, *it is vitally important that regulators stay the course in order to reduce regulatory asymmetries*. The magnitude of the economic distortions is such that a regulated glide path may be necessary in some Member States.

One promising development that bears watching is the termination rate scheme recently notified by the Swedish NRA.[62] The NRA required the largest incumbent to implement a full system of cost accounting and cost-oriented termination rates. Two other operators were required to provide cost accounting, but merely to charge "reasonable and fair prices", presumably no higher than those of the incumbent. The remaining small operators must charge reasonable and fair prices, but were obliged to provide cost accounting data only upon the regulator's request.[63] U.S. experience suggests that systems of this type can achieve low termination rates while burdening only a few operators with full cost accounting and cost orientation.

4.2 Next steps for the United States

The call termination system in the United States is not engendering a mobile termination problem, but intercarrier compensation regimes are nonetheless under significant stress due to increasing competition, differences in the price of different forms of access, and technological and market convergence.

By 2001, the FCC had recognized that the termination mechanisms had become unwieldy and complex. "These regulations treat different types of carriers and different types of services disparately, even though there may be no significant differences in the costs among carriers or services."[64] At the time, mobile telephony and Internet services were placing significant strains on the system.

---

[61] See http://forum.europa.eu.int/Public/irc/infso/ecctf/library.

[62] See the Commission's comments of 9 June 2004, "Case SE/2004/0050: Call termination on individual public telephone networks provided at a fixed location in Sweden: Comments pursuant to Article 7 (3) of Directive 2002/21/EC", at http://forum.europa.eu.int/irc/DownLoad/kVeUAoJ-mTGtGV2OGE-pBsCwUINUn4c0xyNLZFuh2HJ26CawHjUsPD1q6wVIhaNsLk30u/SG%20Greffe%20%282004%29%20D%20202305.pdf.

[63] The proposed approach of the Irish NRA is somewhat similar. See "Consultation on Remedies – Wholesale voice call termination on individual mobile networks", document 04/62b, 8 June 2004. See also the UK approach in case UK/2003/0003.

[64] FCC, *Unified Intercarrier Compensation Regime NPRM*, §5. In her separate statement regarding this NPRM, Commissioner Susan Ness noted that "…we still have in place today a system under which the amounts, and even the direction, of payments vary depending on whether the carrier routes the traffic to a local carrier, a long-distance carrier, an Internet provider, or a CMRS or paging provider. In an era of convergence of markets and technologies, this patchwork of regimes no longer makes sense. What had been a historical artifact may have become an unsustainable anomaly." Chairman Powell observed that "…the rates for interconnecting with the phone network vary depending on the type of company that is doing the interconnecting. In a competitive environment, this leads to arbitrage and inefficient entry incentives, as companies try to interconnect at the most attractive rates."

A number of economists have suggested that call termination charges under CPNP regimes do more harm than good. In 2001, the FCC published two staff working papers, one by Patrick DeGraba, the other by Jay Atkinson and Christopher Barnekov. Both papers argued for elimination of call termination fees, and a migration to a bill-and-keep (i.e. zero fee) regime.[65]

DeGraba noted that U.S. mobile operators typically negotiate bill-and-keep arrangements among themselves, and that these arrangements appear to work well with no significant regulatory oversight.

Analogously, Internet service providers who "peer", or exchange traffic for their respective customers, often do so on a bill-and-keep basis. Laffont, Marcus, Rey and Tirole have noted that, in economics terms, call termination differs from charges associated with Internet peering primarily as a result of the "missing price": receivers do not pay for receiving calls.[66] Thus, the economics of Internet interconnection may provide valuable insights in regard to call termination.

DeGraba also argues that the recipient of a telephone call derives some benefit from that call, and should consequently share in the price of the call.

The FCC issued an NPRM in 2001 in which they proposed to radically simplify the system by migrating to a bill-and-keep regime.[67] The FCC has not ruled on this NPRM.

Today, IP telephony is placing new strains on the call termination system. The FCC has once again raised the question of how best to deal with call termination fees, this time in the IP Enabled Services NPRM.[68]

4.3 Concluding Remarks

The U.S. call termination system has arguably been less problematic than that used in Europe; however, both systems face significant stresses in the years to come. No country has implemented a system that fully and simultaneously avoids regulatory distortions and addresses convergence challenges for all communication services. Further evolution is necessary and inevitable on both sides of the Atlantic.

---

[65] See Federal Communications Commission (FCC) Office of Strategic Planning and Policy Analysis (OSP) Working Papers 33 and 34: Patrick DeGraba, "Bill and Keep at the Central Office As the Efficient Interconnection Regime"; and by Jay M. Atkinson, Christopher C. Barnekov, "A Competitively Neutral Approach to Network Interconnection", both December 2000. Both are available at http://www.fcc.gov/osp/workingp.html.

[66] Jean-Jacques Laffont, J. Scott Marcus, Patrick Rey, and Jean Tirole, IDE-I, Toulouse, "Internet interconnection and the off-net-cost pricing principle", *RAND Journal of Economics*, Vol. 34, No. 2, Summer 2003. An earlier version of the paper is available at http://www.idei.asso.fr/Commun/Articles/Rey/internet.pdf. "Finally, let us compare Proposition 1 with the results in Laffont, Rey, and Tirole (1998a) and Armstrong (1998) for interconnection of telephone networks. A key difference with this telecommunications literature is that in the latter there is a missing price: receivers do not pay for receiving calls… In sum, the missing payment affects the backbones' perceived costs, and it reallocates costs between origination and reception."

[67] FCC, *Unified Intercarrier Compensation NPRM*.

[68] FCC, *In the Matter of IP-Enabled Services*, WC Docket No. 04-36, released March 10, 2004.

The differences in approach and philosophy are significant. Europe is on a path that may lead to more intensive regulation of call termination; the United States is likely to continue on its generally deregulatory trajectory. These differences are largely a function of path dependencies – Europe and America are starting from somewhat different points today.

As we move forward, there is great value to policy experts and practitioners on both sides of the Atlantic in developing a comprehensive understanding of the strengths and weaknesses of both systems. This paper has sought to contribute to that understanding.

Acknowledgments

First and foremost, I would like to thank my FCC colleagues Donald Stockdale and William Sharkey. They contributed a wealth of ideas and insights.

I am grateful to Jane Mago and the management team at the FCC for enabling me to research this topic; however, the opinions expressed are not necessarily those of the FCC.

Special thanks are also due to my outstanding colleagues at the European Commission, including Peter Scott, Richard Cawley, Patrick Kenny, and Sandra Keegan. They gave of their time and knowledge with great generosity.

I am also deeply indebted to the German Marshall Fund of the United States for their support of this study.

National ministry and regulatory colleagues provided additional valuable insights, including Gabrielle Gauthey, Benoît Loutrel and Anne Lenfant of the French ART; Philipp Braun, Annegret Groebel, Cara Schwarz-Schilling and their colleagues at the German RegTP; and Peter Knauth and and Frank Krueger of the German Bundesministerium für Wirtschaft und Arbeit.

Malcolm Harbour, an immensely knowledgeable Member of the European Parliament (MEP), provided useful pointers and guidance.

Sam Paltridge and Dimitri Ypsilanti of the OECD provided many helpful comments.

Industry (and their advisers and counsel) provided important complementary insights. Particularly noteworthy are Michael Bartholomew of ETNO, Andreas Tegge of Deutsche Telekom, Wolfgang Kopf, Rui Pereira, and Volker Stapper of T-Mobile, Carsten Hess of MCI, Wilfried Stratil of Telekom Austria, Christian Duvernoy of Wilmer Cutler, Jim Venit of Skadden Arp, and Winston Maxwell of Hogan and Hartson. Mark Cardwell of Sanford Bernstein provided a wealth of valuable ideas and information. Last but not least, there is Frédéric Donck or Belgacom, whose insightfulness is exceeded only by his grace on the dance floor.

Finally, Jean Tirole and Patrick Rey of the Institut d'Economie Industrielle (IDEI) provided valuable guidance. Indeed, it was a question from Jean that inspired this paper.

I owe a great debt to all of these individuals and organizations, but the opinions are my own, and any errors or omissions are entirely my own responsibility.

Bibliography

Carter, Michael and Wright, Julian, "Interconnection in Network Industries", Review of Industrial Organization 14: 1–25, 1999.

Cellular Telecommunications and Internet Association, *Semi-Annual Wireless Industry Survey* (see http://www.wow-com.com/industry/stats/surveys/).

Crandall, Robert W.; and Sidak, J. Gregory, "Should Regulators Set Rates to Terminate Calls on Mobile Networks?" forthcoming in Yale Journal on Regulation, 2004.

Dewenter, Ralf; and Haucap, Justus, "Mobile Termination with Asymmetric Networks", October 2003, available via SSRN. See also 4th ZEW Conference on the Economics of Information and Communication Technologies, Mannheim, Germany, July 2004.

Doyle, Chris and Smith, Jennifer C., "Market Structure In Mobile Telecoms: Qualified Indirect Access And The Receiver Pays Principle", May 1999.

European Commission, *Directive 2002/19/EC of the European Parliament and of the Council of 7 March 2002 on access to, and interconnection of, electronic communications networks and associated facilities (Access Directive*), Official Journal of the European Communities, L 108, April 24, 2002.

European Commission, *Directive 2002/20/EC of the European Parliament and of the Council of 7 March 2002 on the authorisation of electronic communications networks and services (Authorisation Directive*), Official Journal of the European Communities, L 108, April 24, 2002.

European Commission, *Directive 2002/21/EC of the European Parliament and of the Council of 7 March 2002 on a common regulatory framework for electronic communications networks and services (Framework Directive*), Official Journal of the European Communities, L 108, April 24, 2002.

European Commission, *Directive 2002/22/EC of the European Parliament and of the Council of 7 March 2002 on universal service and users' rights relating to electronic communications networks and services (Universal Service Directive*), Official Journal of the European Communities, L 108, April 24, 2002.

European Commission, *Commission Recommendation of 11 February 2003 on relevant product and service markets within the electronic communications sector susceptible to ex ante regulation in accordance with Directive 2002/21/EC of the European Parliament and of the Council on a common regulatory framework for electronic communications networks and services*, Official Journal of the European Communities, L 114, (2003/311/EC), May 8, 2003.

European Commission, *Commission guidelines on market analysis and the assessment of significant market power under the Community Regulatory Framework for Electronic Communications Networks and services*, Official Journal of the European Communities, C 165, (2002/C 165/03), July 7, 2002.

European Commission, *Ninth report on the implementation of the EU electronic communications regulatory package*, COM(2003) 715 final, 19 November 2003

FCC, *In the Matter of Access Charge Reform, Price Cap Performance Review for Local Exchange Carriers, Low-Volume Long-Distance Users, Federal-State Joint Board On Universal Service* ("CALLS Order"), CC Dockets 96-262, 94-1, 99-249, 96-45, released May 31, 2000.

FCC, *In the Matter of developing a Unified Intercarrier Compensation Regime*, CC Docket 01-92, released April 27, 2001.

FCC, *In the Matter of Implementation of Section 6002(b) of the Omnibus Budget Reconciliation Act of 1993, Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Service*, WT docket no. 02-379, released July 14, 2003.

FCC, *In the Matters AT&T Corp., Complainant, versus Business Telecom, Inc., Defendant. Sprint Communications Company, L.P., Complainant, Business Telecom, Inc., Defendant*, Memorandum Opinion And Order, EB Docket EB-01-MD-001, Released: May 30, 2001.

FCC, *In the Matter of IP-Enabled Services*, WC Docket No. 04-36, Released: March 10, 2004.

FCC Office of Strategic Planning and Policy Analysis (OSP) Working Paper 33: Patrick DeGraba, "Bill and Keep at the Central Office As the Efficient Interconnection Regime", December 2000, available at http://www.fcc.gov/osp/workingp.html.

FCC Office of Strategic Planning and Policy Analysis (OSP) Working Paper 34: Jay M. Atkinson, Christopher C. Barnekov, "A Competitively Neutral Approach to Network Interconnection", December 2000. Available at http://www.fcc.gov/osp/workingp.html.

Laffont, Jean-Jacques; Marcus, J. Scott; Rey, Patrick; and Tirole, Jean; IDE-I, Toulouse, "Internet interconnection and the off-net-cost pricing principle", *RAND Journal of Economics*, Vol. 34, No. 2, Summer 2003. An earlier version of the paper is available at http://www.idei.asso.fr/Commun/Articles/Rey/internet.pdf.

Jeon, Doh-Shin; Laffont, Jean-Jacques; and Tirole, Jean, "On the receiver pays principle", to appear in the RAND Journal of Economics, 2004.

Laffont, Jean-Jacques; and Tirole, Jean, *Competition in Telecommunications*, MIT Press, 2000.

Laffont, Jean-Jacques; Rey, Patrick; and Tirole, Jean, "Network Competition: I. Overview and Nondiscriminatory Pricing" (1998a), *Rand Journal of Economics*, 29:1-37.

Laffont, Jean-Jacques; Rey, Patrick; and Tirole, Jean, "Network Competition: II. Price Discrimination" (1998b), *Rand Journal of Economics*, 29:38-56.

Littlechild, Stephen C., "Mobile Termination Charges: Calling Party Pays vs Receiving Party Pays", available at http://www.econ.cam.ac.uk/dae/repec/cam/pdf/cwpe0426.pdf.

OECD, *OECD Communications Outlook*, 2003.

Exhibit F

INTERNATIONAL TELECOMMUNICATION UNION

ITU WORKSHOP ON
What rules for IP-enabled NGNs?

Document: NGN/02
23 March 2006

Geneva, 23-24 March 2006

INTERCONNECTION IN AN NGN ENVIRONMENT

BACKGROUND PAPER

Advanced draft to be presented for comments

© ITU
March 23, 2006

NOTE

This draft paper has been prepared by J. Scott Marcus (Senior Consultant, WIK Consult, Wissenschaftliches Institut für Kommunikationsforschung, Germany, <S.Marcus@wik.org >) to be presented for comments at the ITU New Initiatives Programme workshop on "What rules for an IP-enabled Next Generation Networks?" held on 23-24 March 2006 at the ITU Headquarters, Geneva. The final version of the paper reflecting the comments will be made available at the event's web site in April 2006. The views expressed in this paper are those of the authors, and do not necessarily reflect those of the ITU or its membership.

This paper, together with the others relevant for NGN debate and prepared under ITU New Initiatives Programme, can be found at http://www.itu.int/osg/spu/ngn/event-march-2006.phtml. The New Initiatives Project on "What rules for IP-enabled NGNs?" is managed by Jaroslaw Ponder <jaroslaw.ponder@itu.int> under the direction of Rober Shaw <robert.shaw@itu.int>.

ACKNOWLEDGEMENTS

I have had the good fortune to work with many outstanding professionals. The engineers at GTE Internetworking/Genuity were superb, and taught me a great deal about interconnection in the IP-based world. I have also benefited enormously from my associations with prominent economists, including Jean Tirole, Patrick Rey, the late Jean-Jacques Laffont, Donald Stockdale, William Sharkey, Simon Wilkie, David Sappington, Patrick de Graba, Stephen Littlechild, and Justus Haucap. For the paper at hand, my WIK colleague Dieter Elixmann provided a very thoughtful and careful review, and Robert Shaw and Jaroslaw Ponder of ITU's SPU provided helpful guidance and feedback.

TABLE OF CONTENTS

page

1 Chapter One: Introduction 5

- **1.1** The migration to IP-based Next Generation Networks (NGNs) 5
- **1.2** To regulate, or not to regulate? 5
- **1.3** NGN core, NGN access 6
- **1.4** A word about the author 6
- **1.5** A road map to the balance of the report 6

2 Underlying Economic Principles 7

- **2.1** The PSTN at the Retail Level 7
- **2.2** The PSTN at the Wholesale Level 8
- **2.3** Retail prices, subsidies, adoption, and utilization 10
- **2.4** The Internet 13
- **2.5** Internet interconnection and PSTN interconnection 15

3 Quality of service 16

- **3.1** The economics of service differentiation and price discrimination 16
- **3.2** Technological considerations for IP/QoS 17
- **3.3** Network externalities, transaction costs, and the initial adoption "hump" 20
- **3.4** Prospects for inter-provider QoS in an NGN world 21

4 Market power and NGN interconnection 22

- **4.1** Sources of market power 22
- **4.2** Addressing market power 23
- **4.3** Remedies for market power, or a "regulatory holiday"? 23
- **4.4** The "network neutrality" debate 25

5 Universal service and NGN interconnection 26

- **5.1** Network externalities, economic distortions, and consumer welfare 27
- **5.2** Intercarrier compensation as a funding mechanism for ICT development 27
- **5.3** Traffic imbalance – the "Robin Hood" effect 27
- **5.4** Policy implications 28

6 Billing and accounting in an IP-based world 28

- **6.1** Protocol layering, services, and the underlying network 28
- **6.2** Point-to-point versus end-to-end measurement 29
- **6.3** Reconciliation of statistics 30
- **6.4** Accounting for Quality of Service 30
- **6.5** Gaming the system 31

7 A Hypothetical Scenario: Interconnection in an NGN world 31

- **7.1** The scenario 32
- **7.2** Regulatory implications for last mile access 32
- **7.3** Regulatory implications for interconnection 32
- **7.4** Peering versus transit 33
- **7.5** Network provider versus application servce provider 35
- **7.6** Implications for differentiated Quality of Service 36
- **7.7** Policy implications 36

TABLES

Table 2.1: Revenue per minute versus monthly minutes of use for mobile services. 11

FIGURES

Figure 2.2: Minutes of use versus revenue per minute for mobile services. 12
Figure 3.1: Packet wait time on a 155 Mbps link.. 19
Figure 7.1: Hypothetical peering arrangements 33
Figure 7.2: Hypothetical peering and transit arrangements.. 35

Executive Summary

This report considers the likely evolution of interconnection arrangements in the context of IP-based *Next Generation Networks (NGNs*).

The NGN represents a synthesis of existing world of the "traditional" *Public Switched Telephone Network (PSTN*) with the world of the Internet. The economic and regulatory arrangements for the two have historically been very different. What should happen when these two worlds collide?

Many of the networks created over the past ten years contain most of the key elements of an NGN. Most, if not all, of the technology necessary for IP-based NGN interconnection has been available for five to ten years. Advanced approaches to interconnection have been slow to deploy, even where the technology has been mature or within hailing distance of maturity.

The NGN interconnection problem is best understood, not as a problem of technology, but rather as a problem in economics. With that in mind, this report seeks to review what is known about interconnection from an economic perspective, in order to reach conclusions about the prospects for deployment going forward and the corresponding implications for policymakers.

A substantial body of economic theory has been developed over the past decade as regards interconnection in the traditional PSTN. A smaller body of solid economic research has emerged in regard to interconnection of IP-based networks. At the level of economic theory, the PSTN and the Internet are *not* worlds apart. Economics provides the necessary bridge between the two worlds, illuminating both the similarities and the differences in these two environments.

This report begins by laying out, for the most part at a non-technical level, the established theory of interconnection, for both the PSTN and the Internet. Wholesale and retail arrangements are considered separately. Most of the observed behavior of these economic networks can be explained in terms of a constellation of known economic effects: market power, the termination monopoly, demand elasticity, network exernalities, transaction costs, service differentiation, price discrimination, and the Coase theorem (which says that private parties can often negotiate arrangements more efficiently than government regulators, provided that necessary preconditions have been met).

With this theory in hand, the report considers the implications for the deployment of differentiated Quality of Service, and of universal service. We also consider the implications of IP-based technology – with the layering, and the changes in industry structure that it implies – service providers become independent of the network, but neither is well equipped to measure or to charge for the other's resource consumption.

The last section of the report represents a hypothetical scenario, a "thought experiment", where the historic wired and mobile incumbent of European country upgrades its networks to an IP-based NGN. We consider the likely results in terms of regulation of the access network, and of interconnection; likely domestic and international interconnection arrangements; and the implications for ubiquitous support of QoS. Key findings include:

- Provided that markets for Internet transit and for consumer broadband Internet access are effectively competitive, a "Coasian" interconnection regime is likely to be more efficient, and more consistent with consumer welfare, than a regulated regime.
- Conversely, where these markets are not effectively competitive, mandates for interconnection at the IP level may prove to be unavoidable, particularly once existing PSTN interconnection is withdrawn. The migration to NGN potentially creates new sources of market power, at the same time that it creates new possibilities for competition.
- Policymakers might consequently be well advised to focus their attention first on ensuring competitive markets, and only secondarily on interconnection.

Current *Calling Party's Network Pays (CPNP*) arrangements contain a number of implicit subsidies. In the world of the NGN, where services providers and networks operators may be different entities, these subsidies need major re-thinking – call termination payments that were intended to finance the terminating network would, by default, flow to independent VoIP service provides who have no network to support. In the absence of termination fees, independent VoIP providers would tend to compete price levels for telephony service, independent of the network, down to levels not greatly above cost, which would appear to be a societally desirable outcome.

The thought experiment does not flatly preclude the possibility that governments might somehow erect a new system of subsidies to replace the old, but it suggests that any subsidy system will be difficult to sustain over time in the face of new forms of competition enabled by the IP-based NGN – all provided, once again, that underlying markets (especially for wholesale Internet transit and for retail Internet broadband access) remain effectively competitive. A system of Coasian private arrangements, in the absence of vertically integrated competitive bottlenecks, seems likely to lead to unsubsidized arrangements at wholesale and retail price levels not greatly in excess of cost.

1 INTRODUCTION

The English novelist Charles Dickens has a series of ghosts show his miserly and misanthropic protagonist, Scrooge, his past, his present, and a grim future. The chastened Scrooge then asks, "Are these the shadows of the things that Will be, or are they shadows of things that May be, only?"[1]

This report considers the problem of network interconnection in the emerging world of the IP-based NGN from the perspective of established economic theory, and then attempts to "paint a picture" of what might happen if the primary wired and wireless incumbent in a major European country were to migrate rapidly and comprehensively to an IP-based NGN in the near future. It is hoped that this thought experiment sheds light on the likely evolution of interconnection in the evolving NGN world; at the same time, it is important to remember that it depicts *one possible* future, hopefully a plausible future, but not necessarily *the* future.

1.1 The migration to IP-based Next Generation Networks (NGNs)

The global electronic communications industry is experiencing something of a "sea change" as it is integrated to an increasing degree with IP-based services. The plans of British Telecom (BT) to replace outright large parts of its existing over the next few years with a 21st Century Network (21CN) are perhaps the most dramatic example,[2] but the same trend is proceeding, perhaps more quietly, in every developed country. In North America, there is less of the rhetoric of the NGN, but much of the same substance.

1.2 To regulate, or not to regulate?

This migration raises many thorny regulatory questions, especially in the area of network interconnection. The *Public Switched Telephone Network (PSTN*), the existing telephony network, operates under a well established set of interconnection rules that have been more than a century in the making. In the Internet, by contrast, interconnection is generally a matter of private bilateral agreements, usually with no regulatory intervention at all. Both systems seem to work reasonably well most of the time in their respective domains, but how should they be combined?

Inevitably, there have been calls to withdraw regulation altogether. As the number of technical alternatives increases, and competition progressively expands, the regulation of electronic communications should wither away altogether.

In the long run, this is probably the right view. Regulatory best practice argues for withdrawal of regulation once markets have become effectively competitive.

But the long run view may not be the most relevant view. As the English economist John Maynard Keynes remarked, "In the long run, we're all dead." This report focuses on events in an intermediate time frame – the next few years, or perhaps at most the next two decades.

Over that time frame, concerns must be raised over complete withdrawal of regulatory obligations in markets where competition is not yet fully effective. The experience of New Zealand, which attempted for years to avoid putting a traditional communications regulator in place, is particularly relevant – their system proved to be unworkable. In fact, the most serious problems were precisely in regard to interconnection, which is the locus of this report. Starting around 2001, they gave it up as a bad job, and implemented lightweight institutions approximating the function of a traditional regulator.[3]

The scenario analysis in this report suggests that the overarching philosophy that the U.K. regulator, Ofcom, has adopted is much more promising: the focusing of reglulation on areas where there are *durable competitive bottlenecks*, enabling competition at the *deepest level feasible*; and the gradual withdrawal of regulation everywhere else.[4]

1.3 NGN core, NGN access

The migration to Next Generation Networks can be viewed as comprising two distinct threads. On the one hand, current PSTN operators are evolving the *core* of their networks so as to use IP-based technology to carry voice traffic, and other applications as well. On the other, many firms are providing increasingly high speed data *access* to the customer premises.

In a recent document,[5] the European Competitive Telecommuncations Association (ECTA) provided definitions that will serve for purposes of this report:

- The first is the deployment of fibre into the local loop, either to the incumbent's street cabinet (+/- max 1km from the customer premises) in conjunction with VDSL(2) deployment or the deployment of fibre all the way to customer premises (typically apartment blocks rather than individual houses). These will be referred to as *access NGNs*.
- The second is the replacement of legacy transmission and switching equipment by IP technology in the core, or backbone, network. This involves changing telephony switches and installing routers and Voice over IP equipment. These will be referred to as *core NGNs*.

These two threads have somewhat different regulatory implications. In this report, our primary focus in on the NGN core. The adoption of broadband access is very much relevant to this migration, and in this sense the migration to the access NGN can be viewed in regulatory terms as simply being faster broadband.

1.4 A word about the author

I should also say a few words about my own background. We all have a tendency to look at issues through the lens of our own experiences. Before starting work at the WIK, a research institute and consulting firm located in Bad Honnef, Germany, I had been the Senior Advisor for Internet Technology at the FCC (U.S.). Prior to that, I was the Chief Technology Officer (CTO) for GTE Internetworking (Genuity, also U.S.), which at the time was one of the largest Internet backbones in the world.

I am well aware that these issues are complex and contentious. My long experience working with the Internet, with the FCC, and generally in North America inevitably predisposes me toward a Bill and Keep intercarrier compensation model; at the same time, I am reasonably well versed in theory and practice in Europe. The perceptive reader will quickly observe that my personal views on these matters do not strictly follow the lines on which these arguments typically proceed. I have attempted to present the issues and the full range of arguments as clearly and as fairly as I could, and to ground my statements clearly in established economic theory and in documented facts. Only the reader can judge how well I have succeeded.

I should add that, while I know something about economics, I do not regard myself as an economist. I am an engineer by training. Nonetheless, I took an economic perspective in this report, because the interconnection challenges with which this report deals are best understood from that perspective.

1.5 A road map to the balance of the report

The next three sections of the report provide general background drawn from economic theory. Section 2 provides interconnection theory, both for the PSTN and for the Internet. Section 3 provides technical and economic background of differentiated service (IP Quality of Service), and of associated price discrimination. Section 4 talks about market power – its sources, its remedies, and its likely evolution in the world of the IP-based NGN. Section 5 is a brief exploration of the relationship between interconnection arrangements and the funding of universal service in an NGN context. Section 6 considers the interaction between interconnection arrangements and interconnection accounting – what can be measured in an IP-based NGN, and how do measurement constraints translate into constraints on what can be charged for? Finally, chapter 7 uses a hypothetical scenario of an NGN migration in Europe to explore how interconnection arrangements might in practice evolve.[6]

2 UNDERLYING ECONOMIC PRINCIPLES

This section provides background on the underlying economics of network interconnection, in order to motivate the discussion that follows. It attempts to present the economics of the PSTN and that of the Internet in an integrated way, and also to provide a consistent view of the various models that have emerged at the retail and at the wholesale levels.

The interconnection of telecommunications networks has been extensively studied in the literature. Many economists would view the authoritative sources as being Laffont, Rey and Tirole (1998a and 1998b),[7] Armstrong (1998),[8] and Laffont and Tirole (2001).[9] I choose to draw primarily on Laffont and Tirole (2001).

The section seeks to provide non-specialists with a non-technical but thorough grounding in the theory and the literature.[10] It also serves to introduce the economics vocabulary that will be used throughout the balance of the paper. Economists may find this section useful primarily to the extent that it provides a comprehensive and integrated view of what is known of interconnection arrangements in the PSTN and in the Internet.

2.1 The PSTN at the Retail Level

Retail arrangements in the world of conventional telephony are, in a sense, familiar to anyone who uses a telephone. Nonetheless, it may be helpful to put them into a broader perspective, in order to provide a comparative context. Most of us live in a single country, and have only limited exposure to alternative arrangements.

2.1.1 Calling Party Pays (CPP) versus Mobile Party Pays (MPP)

In most countries, the party that *originates* (initiates) a call pays a fee for the call, usually as a function of the duration of the call in minutes, and often also as a function of the distance from the originator to the point at which the call *terminates* (is received). In these same countries, the party that receives the call typically is not charged. These arrangements are collectively referred to as *Calling Party Pays (CPP*).

A few countries – notably, the United States and Canada – use an alternative system referred to as *Receiving Party Pays* (*RPP*). Under RPP, the originating party and the terminating party can each be charged by their respective service providers.

In the U.S. and Canada, CPP arrangements are common for fixed line phones, while RPP arrangements are common for mobile phones. For this reason, some experts prefer to refer to these North American arrangements as *Mobile Party Pays* (*MPP*).

In fact, the system in these countries continues to evolve – the most common arrangements today are for plans that are either *flat rate*, or that are flat rate up to some large number of minutes of use (so-called *buckets of minutes* plans).

Each of these systems has its advantages and its disadvantages, and each has adherents and opponents. Both are in need of a major re-thinking as the world evolves to IP-based NGN arrangements.

2.1.2 Cost Causation

CPP calling arrangements have long been the globally most common set of arrangements. They are extremely logical if one starts from the presumption that the party that originated a call presumably wanted the call to complete, and that the originating party can therefore be considered to be both the prime beneficiary and the *cost-causer* of the call.

Analogously, the receiving party has been thought of as a passive party, involuntarily receiving a call from the originator. Again, under this assumption it is natural to refrain from charging the receiving party.

More recently, a number of economists have challenged this view. The American Patrick deGraba has argued that, "… both parties to a call – i.e., the calling party and the called party – generally benefit from a call, and therefore should share the cost of the call." [11]

A recent paper by Doh-Shin Jeon, the late Jean-Jacques Laffont, and Jean Tirole explores the inherent mirror-image relationship between calling and called party, and find that there is no qualitative difference, as "it takes two to tango." In particular, they consider the implications of *receiver sovereignty* – the notion that the receiver always has the option to hang up, and therefore should be viewed as playing an equal or nearly equal role in cost causation.[12]

2.1.3 Usage-based pricing versus flat rate

Consumers appear to have a strong preference for flat rate retail pricing arrangements over usage-based pricing. Flat rate arrangements reduce or eliminate the uncertainty as to what the consumer will have to pay.

Customers tend to respond to flat rate plans by making extensive use of the service in question. In an economic sense, this is a normal and predictable demand elasticity response to a perceived marginal price of zero.

If the marginal usage-based cost to the provider were high, this might lead to inefficient use; however, communications services today are characterized to an ever-increasing degree by significant initial costs and low or very low usage-based marginal costs. Under these circumstances, flat rate plans can be efficient for both the consumer and the provider. The high utilization of the service that flat rate promotes can thus be viewed as a gain in consumer welfare.

The U.S.-based mathematician Andrew Odlyzko has argued that pricing structures will tend to gravitate to flat rate whenever the marginal cost is low enough, and purchases frequent enough: "People react extremely negatively to price discrimination. They also dislike the bother of fine-grained pricing, and are willing to pay extra for simple prices, especially flat-rate ones. …[P]rice discrimination and finegrained pricing are likely to prevail for goods and services that are expensive and bought infrequently. For purchases that are inexpensive and made often, simple pricing is likely to prevail."[13]

Flat rate plans are common in the United States, but much less common outside of North America, largely as a function of differences in the underlying wholesale interconnection arrangements – we return to this point in the following section of this paper. Experience in the U.S. strongly bears out the consumer preference for flat rate services.

For example, AT&T Wireless's offer of Digital One Rate in 1998 provided flat rates across the United States. As long as the mobile customer used not more than some fixed (and possibly large) number of minutes of air time, the customer could place or receive calls to and from any point in the continental United States. The customer would incur no per-minute charges, no long distance charges, and no roaming charges.[14]

Digital One Rate proved to be immensely popular. The success of Digital One Rate effectively forced its mobile competitors to provide a competitive response; however, initially they were hampered by their lack of nationwide scale. The net result was a wave of consolidation, alliances and joint ventures that ultimately resulted in a nationwide market for mobile telephone services with multiple carriers, each offering nationwide plans offering a large bucket of minutes for a flat monthly fee.

Today, flat rate plans are becoming increasingly prevalent in the U.S. for all forms of telephony.[15] As dominant local operators were permitted to offer long distance services, they typically offered flat rate plans with unlimited domestic long distance. IP telephony service providers commonly offer unlimited domestic calls at a flat rate.[16]

Analogously, when America Online introduced flat rate pricing of $19.95 per hour for Internet service in 1996, it resulted in an explosion of consumer adoption – so much so, that the company was hard-pressed to deploy new service quickly enough.

At the level of governmental policy, both the U.S. and the U.K. have implemented measures to enable consumers to avoid per-minute charges when using dial-up to access an ISP.[17] These measures are motivated by the same recognition that true usage-based incremental costs are low, and that the societal value and consumer welfare benefits of increased utilization of the Internet are probably substantial.

2.2 The PSTN at the Wholesale Level

Charging arrangements for the PSTN at the wholesale level mirror the arrangements at the retail level, but only loosely.

The most common arrangement by far is often referred to *calling party's network pays* (*CPNP*). In a CPNP regime, the call receiver's operator assesses some predefined charge per minute to the caller's operator for termination. The call receiver's operator pays nothing.[18] Given that, under a pure CPP retail regime, the receiving party does not pay for the call at all at the retail level, the prevailing view has been that the calling party's *network* should compensate the receiving party's *network* (i.e. the terminating network) for its costs with a payment at the wholesale level.

Bill and Keep, by contrast is a United States term of art that denotes the absence of a regulatory obligation to make payments at the wholesale level. Carriers could conceivably choose to voluntarily negotiate compensation arrangements at the wholesale level, but in general they are not motivated to do so.

Most countries use CPP at the retail level, and CPNP at the wholesale level. Indeed, wherever CPNP is practiced with relatively high per-minute termination fees (e.g. in excess of several cents per minute), the use of CPP at the retail level tends to follow as an economic consequence.

By contrast, only a few countries use Bill and Keep, and they tend to use it selectively. The United States, for example, is CPNP for call to fixed incumbent operators,[19] but is generally effectively Bill and Keep for mobile-to-mobile calls and for calls from one non-incumbent fixed provider to another.[20] France used Bill and Keep for mobile-to-mobile calls until 2004, generally with satisfactory results.

Bill and Keep wholesale arrangements make flat rate retail plans possible, but they do not preclude other arrangements at the retail level.

2.2.1 Calling Party's Network Pays (CPNP) versus Bill and Keep

As has been previously noted, a very extensive literature exists on wholesale call termination arrangements in general.[21] A number of papers specifically address the relative merits of CPNP wholesale arrangements in comparison with Bill and Keep.[22]

There is some tendency in the literature to use the terms CPP and CPNP interchangeably, but this can lead to confusion. For our purposes we define CPNP in terms of *wholesale* payments between operators. CPP, by contrast, relates to *retail* payments from end-users to their operators. CPP and CPNP are often found together, but not always. The wholesale arrangements do not invariably dictate the retail arrangements, nor *vice versa*.

2.2.2 The termination monopoly

CPNP termination leads to a problem that is known as the *termination monopoly*. When you attempt to place a call to someone, you may have a number of choices as to how to originate the call, but in general you have no control over how the call is to be terminated – in general, a single operator is able to terminate calls to any given telephone number. This confers a special form of market power on the terminating operator – hence, the term termination monopoly.

The termination monopoly operates even in markets where competition for call origination is effective, and is by no means limited to large players that have market power on the call origination market. Laffont and Tirole speak of "… the *common fallacy that small players do not have market power and should therefore face no constraint on their termination charges*. … A network operator may have a small market share; yet it is still a monopolist on the calls received by its subscribers. Indeed, under the assumption that retail prices do not discriminate according to where the calls terminate, *the network has more market power, the smaller its market share*; whereas a big operator must account for the impact of its wholesale price on its call inflow through the sensitivity of its rivals' final prices to its wholesale price, a small network faces a very inelastic demand for termination and thus can impose higher markups above the marginal cost of terminating calls."[23]

Consequently, and in the absence of regulation, operators will tend in general to set their termination prices well in excess of marginal cost, and at levels that are also well above those that are societally optimal.[24]

The high termination fees can lead to large economic distortions where regulation is asymmetric. For example, the general practice in Europe prior to 2003 was to limit wired incumbent operators to termination fees based on marginal cost plus a reasonable return on capital; mobile operators, however, generally had unregulated termination rates. This resulted in European mobile termination rates that were an order of magnitude greater than fixed termination rates, and in very substantial subsidization of mobile services by customers of fixed service. A number of economists have argued that these transfer payments constitute an inappropriate subsidy from fixed to mobile services, and a massive economic distortion.[25]

The European Union can be said to generally subscribe to this analysis. Since 2003, the European regulatory framework for electronic communications has in effect treated the termination monopoly as an instance of Significant Market Power (SMP) that national regulators must deal with. In the absence of mitigating factors, all operators – large and small, fixed and mobile – will tend to be assumed to possess SMP. As a result, mobile termination prices have declined somewhat, and are likely to continue to do so in most if not all Member States of the European Union.[26]

Under a Bill and Keep regime, the terminating monopoly problem does not arise. Interconnected operators generally have the opportunity under Bill and Keep to voluntarily negotiate interconnection prices other than zero; however, experience with mobile operators and with non-dominant wired operators (CLECs) in the United States, with[27] mobile operators in France prior to 2004, and with Internet backbones suggests that interconnection prices in the absence of a regulatory mandate will most often be voluntarily set to a price of zero.[28]

2.2.3 The relationship between wholesale intercarrier compensation and retail prices

If traffic is balanced between two operators, and if they were to charge identical termination fees to one another, then there would be no net payment between them. This is true whether the termination fees are low or high. Since termination fees do not change net payments under these conditions, there may be a temptation to think that termination fees do not matter very much.

Laffont and Tirole refer to this as the *bill-and-keep fallacy*. "It is correct that a change in the access charge need not affect the (absence of) net payment between the operators, but the access charge affects each network's perceived marginal cost and therefore retail prices. It is, therefore, *not* neutral even if traffic is balanced."

Each operator views its payments to other operators as a real cost. Other things being equal, operators will tend to be reluctant to offer service at a marginal *price* below their marginal *cost*. For on-net calls – calls from one subscriber of a network to another subscriber of the same network – operators can and often do offer lower prices that correspond to the operator's real costs.[29] For *off-net* calls (calls to a subscriber of another network), however, it is unusual to see retail prices below a "high" wholesale call termination rate,[30] *even where termination payments are likely to net to zero*. This probably reflects the operators' understandable fear of *adverse selection* – if they set their retail price for off-net calls too low, they may attract too many of precisely those users whose calling patterns are such as to cause them to place more off-net calls, thus generating a net payment (an *access deficit*) to other operators.[31]

2.3 Retail prices, subsidies, adoption, and utilization

As we have seen, high termination fees tend to lead to high retail prices for placing calls. (Under CPP retail arrangements, there is no charge for calls that are received, whether termination fees are low or high.) In particular, high call termination rates preclude flat rate or buckets of minutes plans at the retail level. As we might expect, the higher marginal prices at the retail level tend to depress call origination – this is the well-known phenomenon of *demand elasticity* (or the *price elasticity of demand*). As the price of some good or service goes up, we will prefer to purchase less of it if we can.

The American economist Patrick de Graba described these relationships succinctly in a widely read FCC white paper[32]:

> One source of inefficiency is that existing termination charges create an "artificial" per-minute cost structure for carriers that will tend to result in inefficient per-minute retail prices. In unregulated, competitive markets, such as the markets for [mobile telephony] services and Internet access services, retail pricing is moving away from per-minute charges and towards flat charges or two-part tariffs that guarantee a certain number of free minutes. This suggests that few costs are incurred on a per-minute basis, and that flat-rated pricing will lead to more efficient usage of the network. The existing reciprocal compensation scheme, which requires the calling party's network to pay usage sensitive termination charges to the called party's network, imposes an "artificial" per-minute cost structure on

carriers which, if retail rates are unregulated, will likely be passed through to customers in the form of per-minute retail rates. Such usage sensitive rates thus would likely reduce usage of the network below efficient levels.

DeGraba also notes that "…[t]he ISP market illustrates the importance of rate structure on usage. When AOL changed from usage sensitive rates to a flat charge for unlimited usage in late 1996 the number of customers and the usage per customer rose dramatically and other competitors soon followed. … Similarly, the introduction by [mobile operators] in the United States of pricing plans that include 'buckets' of minutes appear [sic] to have contributed significantly to the growth in wireless usage."[33]

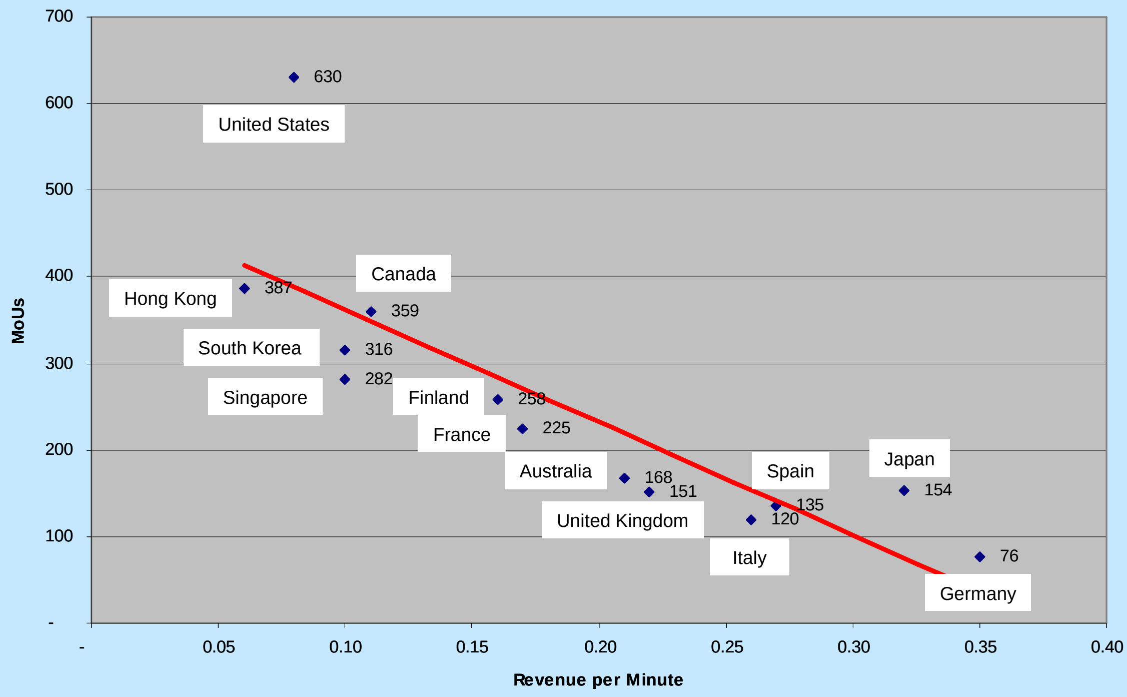
The relationship between termination fees, retail prices, and usage of the service by consumers can more readily be appreciated in regard to the mobile sector, since termination fees and in some cases retail prices are often regulated for fixed incumbents.[34] The investment firm Merrill-Lynch provides an annual analysis of the mobile sector in a number of countries that the U.S. FCC routinely quotes in their annual reports on competition in the U.S. mobile industry,[35] and that other economists also find it convenient to quote.[36] This data is shown in Figure x. For this purpose, we can take the *revenue per minute* for all carriers in a country as being a reasonable proxy for retail price, and a proxy that avoids the complexity of dealing with a plethora of different pricing plans and promotional offers. The *minutes of use* includes minutes of both origination and termination, whether charged or not. Based on this data, Figure 2.1 below depicts the relationship between revenue per minute and minute of use for a number of countries.

Table 2.1: Revenue per minute versus monthly minutes of use for mobile services.

Country	Revenue per Minute ($)	Minutes of Use
USA	0.08	630
Hong Kong	0.06	387
Canada	0.11	359
South Korea	0.10	316
Singapore	0.10	282
Finland	0.16	258
France	0.17	225
Australia	0.21	168
Japan	0.32	154
UK	0.22	151
Spain	0.27	135
Italy	0.26	120
Germany	0.35	76

Source: *FCC*, Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, 10th Report (10th CMRS Competition Report*), July 2005, Table 10, based on Glen Campbell et al.*, Global Wireless Matrix 4Q04, *Global Securities Research & Economics Group, Merrill Lynch, Apr. 13, 2005.*

Figure 2.2: Minutes of use versus revenue per minute for mobile services.



Source: The data derive from FCC, *Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, 10th Report (10th CMRS Competition Report*), July 2005, Table 10, based on Glen Campbell et al., *Global Wireless Matrix 4Q04*, Global Securities Research & Economics Group, Merrill Lynch, Apr. 13, 2005.

The data clearly suggest that lower retail prices will tend to be associated with significantly higher utilization, expressed in minutes of use per month, and *vice versa*. The United States – with per-minute revenues of just $0.08 per minute, but with a marginal price that many users perceive (somewhat inexactly)[37] as zero – experiences more than eight times as much consumption, expressed in terms of minutes used per month, as a country like Germany, where average revenue per month is about $0.35 per month.

Strictly speaking, what is depicted is not demand elasticity – these are not the same customers, and the mobile services that they are using are not mutually substitutable, because they exist in different countries. But the data strongly suggest that demand is elastic, which is to say that a lower price will lead to notably higher utilization.

Thus, Bill and Keep arrangements make possible retail plans with flat or bucketed rates that are perceived as having zero marginal price, and that consequently generate heavy and efficient usage; however, these same plans tend to be associated with slower adoption of mobile services by consumers. The more common CPP/CPNP arrangements generate effective subsidies to mobile operators. Portions of these subsidies are returned to consumers[38] in the form of low or zero commitment periods, subsidies on handset purchase, and low or zero fixed (monthly) fees. CPP/CPNP systems also may be more hospitable to pre-paid arrangements than are Bill and Keep arrangements.

The low fixed fees and low monthly price make it very easy for a consumer to procure a new mobile service. The consumer need make only a small initial investment and commitment. To the extent that the consumer intends primarily to receive calls, rather than to originate them, the total cost will remain low. Conversely, the operator benefits from termination fees in excess of marginal cost whenever the consumer receives calls. The low, subsidised initial price is a clear case of "giving away a razor in order to sell the blades".

The combined effect is to encourage consumers to initially adopt mobile service.[39]

In Europe, there is a growing sense that it is no longer necessary to subsidize the adoption of mobile services.[40] A number of European countries have penetration rates in excess of 100%.[41] Conversely, Crandall and Sidak argue persuasively that mobile phone penetration in the United States (currently at 65%, and growing by about five points per year) is within just a year or two of reaching European levels, and that Canada is following the same pattern but trailing by a few years.[42] Thus, countries that have buckets of minutes arrangements, based on Bill and Keep wholesale arrangements, tend to experience slower take-up, but can in time achieve reasonably high adoption rates.

In particular, these termination arrangements effectively subsidize mobile operators at the expense of fixed operators and fixed customers. This subsidy is arguably irrational and inappropriate.

To re-cap, what appears to be known is:

- Bill and Keep wholesale arrangements enable low or zero retail per-minute usage fees, but higher initial and fixed per-month fees;
- CPNP wholesale arrangements tend conversely to preclude flat rate or buckets of minutes retail arrangements, leading instead to low initial and per-month fees but high per-minute usage fees;
- Countries with buckets of minutes retail arrangements tend to experience high and efficient utilization, but slower adoption of mobile services;
- Countries with conventional CPNP/CPP arrangements tend to experience lower utilization, but faster adoption of mobile services.

An obvious implication is that countries where the market for mobile services is already mature or saturated might want to consider changing to Bill and Keep arrangements. Conversely, developing countries anxious to foster the widespread initial adoption of mobile services might prefer CPP/CPNP.

2.4 The Internet

2.4.1 Peering versus Transit

The two most prevalent forms of Internet interconnection are *peering* and *transit.* For a definition of these terms, we turn to a publication of the Network Reliability and Interoperability Council (NRIC), an industry advisory panel to the U.S. FCC:

> *Peering* is an agreement between ISPs to carry traffic for each other and for their respective customers. Peering does not include the obligation to carry traffic to third parties. Peering is usually a bilateral business and technical arrangement, where two providers agree to accept traffic from one another, and from one another's customers (and thus from their customers' customers). …
>
> *Transit* is an agreement where an ISP agrees to carry traffic on behalf of another ISP or end user. In most cases transit will include an obligation to carry traffic to third parties. Transit is usually a bilateral business and technical arrangement, where one provider (the transit provider) agrees to carry traffic to third parties on behalf of another provider or an end user (the customer). In most cases, the transit provider carries traffic to and from its other customers, and to and from every destination on the Internet, as part of the transit arrangement. In a transit agreement, the ISP often also provides ancillary services, such as Service Level Agreements, installation support, local telecom provisioning, and Network Operations Center (NOC) support.
>
> Peering thus offers a provider access only to a single provider's customers. Transit, by contrast, usually provides access at a predictable price to the entire Internet. … Historically, peering has often been done on a bill-and-keep basis, without cash payments. Peering where there is no explicit exchange of money between parties, and where each party supports part of the cost of the interconnect, … is typically used where both parties perceive a roughly equal exchange of value. Peering therefore is fundamentally a barter relationship.[43]

In the literature, there is some tendency to assume that peering is invariably free, but this is not necessarily the case. Peering is a technical rather than an economic matter; the economic consequences then follow. When the author was in charge of peering policy for GTE Internetworking (at the time one of the five largest

Internet backbones in the world), about 10% of our peering relationships involved payment. These payments were not a function of the relative sizes of the participants; rather, they were a reflection of traffic imbalance. For Internet backbones interconnected at multiple points by means of shortest exit routing, the traffic *received* from another network must on the average be carried further, and must therefore cost more, than the traffic *sent* to the other network ; consequently, when traffic is unbalanced, the network that *sends* more traffic incurs lower cost than the network that *receives* more traffic.[44]

2.4.2 Roughly hierarchical structure

It is impractical for every ISP to directly peer with every other ISP.

A few years ago, *Boardwatch Magazine* listed more than 7,000 ISPs in the United States alone.[45] I am aware of no current reliable data on the number of distinct ISPs in the world, but the number of *Autnonomous System Numbers (ASNs*) currently assigned sets an effective upper limit, since it represents the maximum number of distinct networks that could be using BGP routing to exchange IP data. According to data maintained by the IANA, the responsible global assignment authority, this number might be somewhere between 30,000 and 40,000 networks.[46]

A few years ago, the author was in charge of peering policy for one of the largest Internet backbones in the world at the time. As of 2001, we had perhaps 50 peering relationships. At the same, my staff felt that technical constraints would limit the firm to perhaps a couple of hundred peering relationships at the maximum.

Aside from any remaining technical constraints, the number of peering relationships will in practice also be limited by:

- The costs of providing connections to each of a large number of peering partners; and
- The significant administrative costs associated with maintaining peering agreements with a large number of organizations.

For all of these reasons, the maximum number of peers that an organization could cost-effectively accommodate is perhaps two orders of magnitude less than the number of independent IP-based networks in the world.

This is why the system that has evolved uses a combination of peering and transit relationships to connect to all Internet endpoints in the world. In practice, the Internet can be viewed as a very roughly hierarchical system, comprising (1) a very few large providers that are so richly interconnected as to have no need of a transit provider, and (2) a much larger number of providers who may selectively use peering with a more limited number of partners, and use one of more transit providers to reach the destinations that their peering relationships cannot.[47]

Milgrom et. al. analyzed these peering and transit relationships in depth. Their "… economic analysis of Internet interconnection concludes that routing costs are lower in a hierarchy in which a relatively small number of core ISPs interconnect with each other to provide full routing service to themselves and to non-core ISPs."[48]

2.4.3 Incentives to interconnect

A body of economic theory that first appeared twenty years ago analyzed incentives of firms to conform standards when participating in markets characterized by strong network externalities.[49] Economic analysis suggested that a firm that had a large or dominant customer base would not wish to adhere perfectly to open standards, because full adherence (and thus full fungibility with competing products or services) would limit the ability of the dominant firm to exploit its market power. Some years later, it was recognized that substantially the same analysis applied to network interconnection.

The issue came up in the context of a number of major mergers, and was analyzed at length in Cremer et. al.[50] Again, the conclusion was that, in a market for Internet backbone services characterized by strong network externality effects, if one backbone were to achieve a very large share of the customer base, it would have both the ability and the incentive to disadvantage its competitors. Conversely, as long as the largest backbone had not too large a share of the customer base, and as long as the disparity between the largest backbone and its nearest competitors were not too great, incentives to achieve excellent interconnection would predominate.

Milgrom et. al. studied backbone peering and reached similar conclusions: "A simple bargaining model of peering arrangements suggests that so long as there is a sufficient number of core ISPs of roughly comparable size that compete vigorously for market share in order to maintain their bill-and-keep interconnection arrangements, the prices of transit and Internet service to end users will be close to cost."[51]

The thresholds at which the potential anticompetitive effects might dominate have not been rigorously determined.[52] What can be said today is that Internet interconnectivity is near perfect, and that peering disputes are, in a relative sense, quite rare. It is reasonable, based on these indicia, to conclude that the global Internet is operating well below the thresholds where the anticompetitive effects would predominate.

2.5 Internet interconnection and PSTN interconnection

In this section, we seek to compare and contrast interconnection in the PSTN world with peering in the world of the Internet. First, we briefly review some results from economic theory. Second, we consider the significance of the absence, in general, of regulation of Internet peering. Third, we draw parallels between the largely unregulated mobile telephony sector in the U.S. and the Internet.

2.5.1 Economic theory and the "missing payment"

Interconnection in the world of the Internet evolved independently from interconnection in the PSTN. There is some tendency, due in part to differences of culture and orientation of the respective market participants, to assume that these are different worlds, with little or no commonality.

In fact, the economic models for intercarrier compensation in the two worlds are closely linked. The definitive works on intercarrier compensation in the world of the PSTN are generally considered to be Armstrong (1998)[53] and Laffont, Rey and Tirole (1998a)[54]. In Laffont et. al. (2005)[55], we compared Internet backbone peering with these economic analyses of the PSTN and found:

> A key difference with this telecommunications literature is that in the latter there is a missing price: receivers do not pay for receiving calls … The missing price has two important implications:
>
> **Pricing.** The operators' optimal usage price reflects their perceived marginal cost. Comparing the two perceived marginal costs of outgoing traffic with and without receiver charge, for given access charge and market shares, the price for sending traffic is higher (lower) than in the presence of reception charges if and only if there is a termination discount (markup). … In sum, the missing payment affects the backbones' perceived costs, and it reallocates costs between origination and reception.
>
> **Stability in competition**. When networks are close substitutes, and receivers are not charged, there exists no equilibrium unless the access charge is near the termination cost.

2.5.2 The unregulated Internet

An important difference between PSTN interconnection and Internet interconnection is that the latter has generally not been subject to regulation. Bilateral negotiations for Internet interconnection have in most cases led to very satisfactory arrangements for all parties concerned.[56] This outcome is best understood in terms of (1) the Coase Theorem, and (2) issues of market power.

The Nobel-prize-winning economist Ronald H. Coase has argued, most notably in a famous 1959 paper,[57] that private parties could in many cases negotiate arrangements to reflect economic values far more accurately and effectively than regulators, provided that relevant property-like rights were sufficiently well defined. The generally positive experience with Internet peering appears to bear this out.

If one party to a bilateral negotiation had significant market power, and the other lacked countervailing power, then one might expect that the Coasian negotiation might either break down or might arrive at an outcome that was not societally optimal. In general, this does not appear to be the case at present. To date, it has been widely if not universally recognized that Internet backbones do not possess significant market power.

The migration to IP-based NGNs is one of several interrelated trends[58] that have the potential to change this assumption in a number of ways. On the one hand, as wired incumbent telephone companies and, in some countries, cable companies evolve into vertically integrated enterprises that are also significant Internet backbones, it is entirely possible that they might leverage the market power associated with last mile facilities into their Internet role. Whether this is actually the case for a specific firm or a specific country would need to be evaluated based on market developments in that country, and also through the lens of that country's regulatory and institutional arrangements. Some countries are well equipped to deal with market power; others are not.

At the same time, market power may be mitigated by the emergence and deployment of technological alternatives. Broadband Internet over cable television already has some tendency to mitigate the market power of telephone incumbents. To the extent that broadband over powerline, broadband wireless and other alternatives achieve widespread deployment, they could go a long way to ameliorating or preventing the emergence of market power.

All things considered, this author is of the opinion that:

- unregulated, Coasian Internet interconnection arrangements continue to work well today in most cases, but that
- regulators will need to pay *more*, not less, attention to potential problems in this regard for some years to come.

2.5.3 Analogy of Internet peering to US mobile-mobile interconnection

In the United States, mobile operators have generally been under no regulatory obligation to interconnect with one another; nonetheless, privately negotiated Coasian wholesale interconnection arrangements have worked well. The sector has tended to operate on a Bill and Keep basis. [59] Retail pricing arrangements are completely unregulated, but operators and consumers have increasingly chosen flat rate (buckets of minutes) plans.

The parallels to Internet peering are striking. This experience reinforces the notion that the predicted economic outcome, in a market characterized by strong network externalities, a lack of market power, and no regulatory constraints, is (1) for good interconnectivity and interoperability, and (2) for Bill and Keep arrangements. Moreover, this experience reinforces the notion that these results flow from the underlying economics, and not from any unique technological property of the Internet.

3 QUALITY OF SERVICE

The IP-based NGN is envisioned as providing different levels of *Quality of Service (QoS*), each perhaps offered at a different price, in order to support applications such as real time voice and video on the same IP-based multi-purpose network as data.

In this section, we consider the economics of QoS service differentiation, the technical QoS requirements of applications such as real time voice, the implications of network externalities for adoption of QoS service differentiation, and the implications for long term widespread adoption of QoS differentiation.

3.1 The economics of service differentiation and price discrimination

The basic notion of service differentiation is not new,[60] and the underlying economics have been well understood for many years.[61] Service differentiation recognizes that different consumers may have different needs and preferences, which translate in economic terms into a different *surplus* (the difference between perceived benefits and cost) deriving from the purchase of one service versus another. Service providers can choose to offer tailored products that will be preferred only by certain consumers, or not.[62] In practice, they general target their distinct offers at different *groups* of consumers (second order price discrimination) rather than targeting different individual consumers (first order price discrimination).

We experience service and price differentiation every day. We drive into a gas station, and choose to purchase regular gasoline or premium. We purchase a ticket for an airplane or train, and choose to purchase either economy or first class. To the extent that the amenities offered in first class have value to us, they increase our surplus, which in turn increases the price that we are willing to pay. The airline charges a higher price because they recognize that those customers that value the amenities are willing to pay the higher price.

Even though the benefits of service differentiation are obvious, it enjoys only mixed public acceptance in the context of industries that have historically provided *common carriage*. A long-standing tradition, particularly in England and in the United States, is that certain industries should serve the public *indifferently*. This indifference is taken to imply that *price discrimination* is not allowed. It is largely as a result of these attitudes that airline prices, for example, were regulated for many years.

Today, economists would generally agree that deregulation of the airline industry in the United States and elsewhere (which permitted the airlines to price discriminate) has provided greater consumer choice, and prices that are on the average lower than they would have been had the industry remained regulated.[63] Consumers have had to adjust to the fact that the person sitting in the adjacent seat may have paid a much higher, or a much lower price than they did; nonetheless, overall consumer welfare has improved.

The airline experience in the United States demonstrates both the opportunities and the risks associated with price discrimination. As the economist Alfred E. Kahn (both a proponent and a primary implementer of airline deregulation in the U.S.) has observed, competition on many air routes proved to be limited to only one or two carriers. “In such imperfect markets, the major carriers have become extremely sophisticated in practicing price discrimination, which has produced an enormously increased spread between discounted and average fares, on the one side, and full fares, on the other. While that development is almost certainly welfare-enhancing, on balance, it also raises the possibility of monopolistic exploitation of demand-inelastic travelers.”[64] In other words, those consumers with limited flexibility in their travel requirements could be charged a high premium with impunity. In markets with effective competition, service differentiation and associated price discrimination will tend to enhance consumer welfare. In markets characterized by significant market power, price discrimination could detract from consumer welfare. The airline industry in the U.S. represents an intermediate case, characterized by imperfect competition.

Laffont et. al. (2003)[65] provides a fairly detailed analysis of Internet backbone peering from an economic perspective. In it, we considered possible service differentiation in terms of the mean and variance of packet delay, and in terms of network reliability. We assumed distinct costs for sending and receiving traffic, each proportionate to the total volume of traffic, and we also assumed access charges (either symmetric or asymmetric) proportionate to the volume of traffic, but independent of any consideration of distance. Under these assumptions, symmetric access charges lead to stable competition. In the absence of service differentiation, the backbones would tend to compete away their profits; however, service differentiation between networks can enable the backbones to earn a positive profit.

3.2 Technological considerations for IP/QoS

We now turn to the technological underpinnings of differentiated QoS in an IP network. First, we touch briefly on communications protocol issues; then, we consider application requirements as regards the mean and variance of packet delay. With that established, we consider protocol performance, and discuss the implications for the prospects of widespread adoption.

3.2.1 DiffServ, RSVP, MPLS

By the early Nineties, it had already become obvious to the engineering community that real-time bidirectional voice and video communication could potentially benefit from delivery guarantees on delay. This led to a series of standards efforts – first, the *RSVP*-based Integrated Services Architecture, and then to *Differentiated Services (DiffServ)*.

RSVP provided a comprehensive end-to-end QoS management architecture. Over time, it came to be viewed as hopelessly complex,[66] and was effectively abandoned in favor of DiffServ. DiffServ provides a simple means of specifying, on a hop-by-hop basis, the desired performance characterstics – it is then up to the network to meet those requirements as well as it can.

DiffServ should thus be viewed as a *signaling* mechanism. Technically, it is trivial. The implementation of QoS *within* an IP-based network, with or without DiffServ, has been straightforward with or without DiffServ for at least a decade. Implementation of QoS *between or among* independently managed IP-based networks has never gotten off the ground. Given that the technology is fairly simple, the answers clearly lie in business and economic factors.

3.2.2 Application requirements for bounded delay

Some readers might perhaps assume that all voice and video traffic requires assured quality of service; in reality, however, assurances on the mean and variance of delay are required only for services that involve bidirectional (or multidirectional) voice and video in real time.

The receiving application typically implements a jitter buffer that can be used to smooth the variability in end to end delay. For streaming (one way) audio or video, most users will tolerate a delay of a few seconds when the application starts up. After that, a jitter buffer can typically deal with a considerable amount of variable delay.

For real time bidirectional voice and video, however, users will tend to "collide" if the end to end delay exceeds about 150 to 200 milliseconds. They will both start speaking at roughly the same time, because neither can initially discern whether the other is speaking.[67] This imposes a practical ceiling on the delay that the jitter buffer can allow.

3.2.3 Analysis of delay

This delay in turn imposes limits on both the mean and the standard deviation of delay for the traffic. In an IP-based network, the traffic is composed of individual packets. The delay for these packets can be viewed as comprising a fixed component (based primarily on the speed of signal propagation along the path from send to receiver, and thus dependent primarily on the distance along the path, and also on the deterministic delay to "clock" the packet onto each outbound data transmission link) and a variable component (based on queuing delays in each router through which the packet must pass, especially those associated with gaining access to the outbound transmission link). For a given traffic flow, the unidirectional delay can thus be viewed as a probability distribution with a mean and a standard deviation.

The ability to achieve a round trip delay of not more than 150 milliseconds depends on both the mean and the standard deviation of delay. It is a classic statistical confidence interval problem – it is necessary that the "tail" of the distribution in excess of about 150 milliseconds be suitably small. Note that an occasional outlier is generally permissible – as an example, the *codecs* (coder-decoders) used for Voice over IP (VoIP) services typically interpolate over missing data, and the human ear does a surprisingly good job in compensating for very short data losses. Human speech presumably incorporates a great deal of redundant information that can be used to fill in the gaps.

Fixed delay can be viewed as comprising propagation delay (which is a consequence of the large but finite speed of light) and clocking delay (which is a function of the speed of the transmission link).

We often forget that the speed of light *is* a meaningful constraint. In vacuum, light travels about 300 Km in a millisecond. Signal is not quite as fast when propagating through wires or fiber; moreover, transmission paths (e.g. fiber runs) do not proceed in a geometric straight line. For intercontinental calls, propagation delay can consume a significant fraction of the 150 millisecond budget.

Clocking delay is a function of the speed of the transmission link. Over a dial-up connection to the Internet, clocking delay poses a serious constraint. Over broadband media, it is much less of an issue. In the core of the Internet, the links are very fast indeed, so the deterministic clocking is correspondingly small.

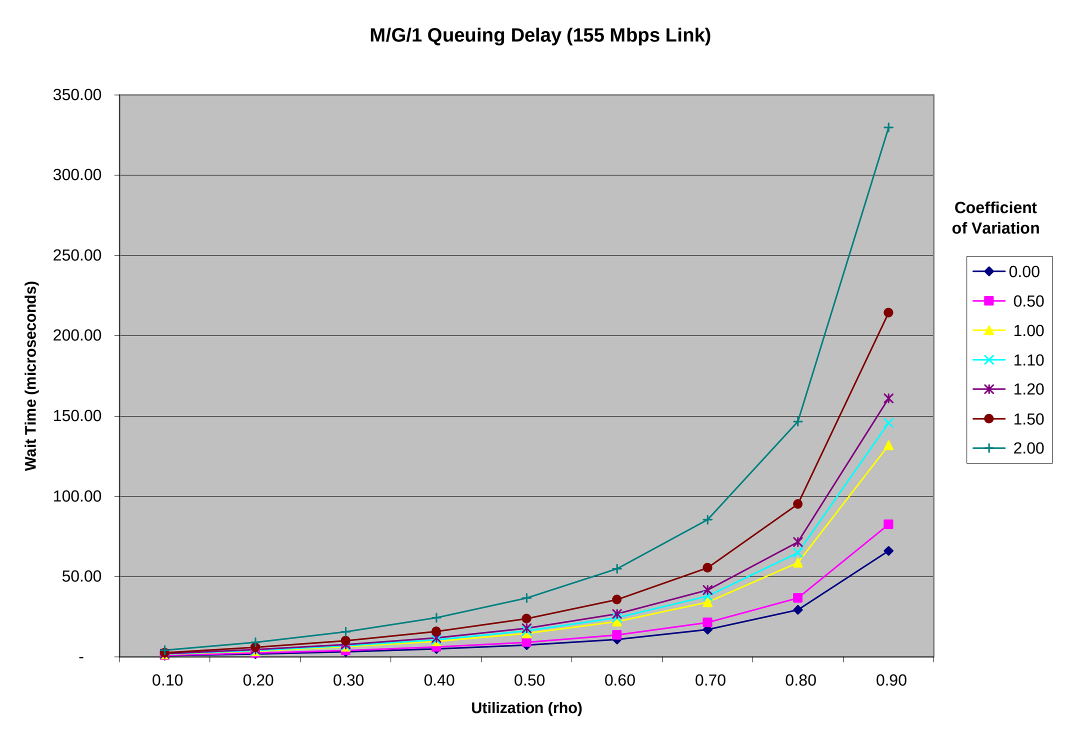
Variable delay is best modeled and analyzed on a hop by hop basis. At each hop, it primarily reflects the queuing delay waiting to clock the traffic onto an outbound link. (Queuing delay for the processor of the router is also possible, but unless the processor is saturated it is generally small enough to ignore.) This variable delay can be analyzed using a branch of mathematics known as *queuing theory* – the science of waiting lines.[68]

Queuing theory tells us that average variable delay reflects three things:

- The average service time (in this case, the deterministic clocking delay);
- The load on the server, which we can think of as the percent of time that it is busy; and
- The variability of the service time, expressed as a coefficient of variation (the standard deviation divided by the mean).

What queuing theory tells us about variable delay in the core of the large IP-based networks is that, *in a properly designed network and under normal operating conditions*, variable delay plays only a very minor role. Figure xxx below depicts the average packet wait time for a 155 Mbps data link, which is the *slowest* link that one would expect to find in the core of a modern Internet backbone.

Figure 3.1: Packet wait time on a 155 Mbps link



Among the family of curves shown, the one corresponding to a coefficient of variation of 1.20 is the one the accords most closely with observational experience around 2001, the most recent date on which this author had access to industry statistics.[69]

The computed average wait time per hop, even at a utilization of 90%, is about 150 *microseconds*. Note that this is *three orders of magnitude* less than the delay budget of 150 *milliseconds*. Beyond this, consider that many backbone links today are one or two orders of magnitude faster than 155 Mbps, with predicted delays correspondingly smaller.

This is not to say that delay could never be a problem. The same queuing theory analysis tells us that, as utilization approaches 100%, predicted mean wait time increases with no upper bound. But no network should be *designed* to operate routinely at those levels. Saturation will occur either as a result of (1) poor planning or forecasting on the part of the network designer, or (2) substantial failures elsewhere in the network that necessitate re-routing of traffic.

3.2.4 Implications for market prospects for QoS

The analysis in the preceding section has significant implications as regards the willingness of customers to pay a surcharge for QoS (in the sense of staticsally bounded delay).

DiffServ-based QoS capabilities cannot speed up a network; they can only prevent it from slowing down (for certain packets) under load. They generally determine (1) which queued packets are served first, and (2) which queued packets are discarded when there is insufficient room to store them.

Under most circumstances, these effects will be too small for the end user to perceive.

It should come as no surprise that end users are unwilling to play a large surcharge for a performance improvement that is not visible to them.[70]

This is not to say that there is no commercial opportunity for inter-provider QoS; rather, it argues that the opportunities will not necessarily be found in the core of the network, which is the place where most people tend to look for them.[71] Instead, QoS will tend to be commercially interesting:

- Within a single provider's network, where the costs of implementation are also low;
- For slower circuits at the edge of the network;
- For shared circuits to the end user (e.g. cable modem services);
- When one or more circuits are saturated;
- When one or more components have failed;
- When a *force majeure* incident (a natural or man-made disaster) has occurred; and especially
- Where more than one of these factors is present.

Providers may also find that offering QoS provides a competitive advantage in attracting new customers, even if those customers are unwilling to pay a large premium.

3.3 Network externalities, transaction costs, and the initial adoption "hump"

The technological capability to deploy differentiated QoS capability at reasonable cost has existed for at least ten years, and has in fact been deployed within many networks. Why has there been so little deployment between or among networks?

The explanation has very little to do with technology, but a great deal to do with economics – specifically, with the economics of *network effects (*or *network externalities)*. An economic market is said to experience network effects when the service becomes more valuable as more people use it. Differentiated QoS is typical of capabilities that take on value only as more networks and more end-users adopt them.

The economist Jeffrey H. Rohlfs has written extensively on the subject of network effects, noting that many new high technology services encounter difficulty in achieving sufficient penetration to get past an initial adoption hump.[72] A certain number of end-users might take up a product or service based solely on its intrinsic value, but that is likely to be far fewer end-users than the number that would take up the service if everybody else did. The market can easily settle into equilibrium at a number of end-users that is far less than the level that would be societally optimal.

The initial adoption hump is often exacerbated by complementarities. A service cannot get launched because it depends on supporting upstream or downstream products and services. CD players could not have succeeded in the marketplace without a substantial inventory of music to play on them. Television sets could not have succeeded without programs to watch. Personal computers could not have succeeded without software to run on them.

Different successful offerings have met this challenge in different ways. In some cases, government intervention has been required. Ubiquitous telephone service is explicitly or implicitly subsidized in many countries – this is referred to as *universal service*. The initial adoption of CD players was facilitated by the fact that the companies that made the players – Phillips and Matsushita – also had interests in studios, could profit on both sides of the market, and were consequently highly motivated to ensure that both players and content were available. The deployment of VCRs in the United States was facilitated by an initial deployment for time shifting of programs – a market for the rental of videos did not emerge until enough devices had worked their way into the hands of consumers.

Certain Internet capabilities have deployed effortlessly – for example, the worldwide web. In many cases, the successful capabilities benefit from the end to end principle – they can be implemented by end-user organizations or consumers, without requiring any action at all on the part of the providers of the underlying IP-based network.

Conversely, other capabilities have tended to deploy at a glacial pace or to stall for reasons not necessarily related to technology, notably including IP version 6 (IPv6), DNS security (DNSSEC), and multicast. A common characteristic among the stalled capabilities is that, rather than being end to end features independent of the network, the stalled capabilities require concerted action and concerted change to the core of the network. Regrettably, inter-provider QoS seems to clearly fit the profile of the stalled capabilities.

Common characteristics among the slow-deploying capabilities include:

- Benefits that are in some sense insufficient: too limited, too difficult to quantify, too slow to appear, too difficult for the investing organizations to internalize.
- Limited benefits until the service is widely deployed.
- The need for coordination among a large number of organizations, leading to high economic *transaction costs* (the cost for a network or an end-user to adopt the service).

If the tangible economic benefits were well in excess of the costs, the services would deploy effortlessly. There are services where the benefits to the organizations that would have to make the investments do not clearly exceed the costs – consequently, the investments are made slowly if at all. The unfavorable relationship between costs and visible benefits hinders initial deployment, and thwarts attempts to reach critical mass and thereby to get beyond the initial adoption hump.[73]

3.4 Prospects for inter-provider QoS in an NGN world

For inter-provider QoS, the benefits in most cases may not be compelling for reasons outlined in Section 3.2 of this paper – in the absence of differentiated QoS, the performance of best efforts traffic will tend to be perfectly adequate in most networks most of the time, and consumers are unlikely to perceive a difference that they are willing to pay for. Moreover, the benefits are limited by the number of other providers that support QoS – the benefits to the first few providers are quite limited.

Conversely, the number of parties that would have to come to agreement to achieve a globally interconnected QoS-capable world is very large.[74] If every pair of providers requires a contractual agreement in order to put QoS in place, then a world with thousands of independent providers will require literally millions of agreements – and complicated agreements at that, for reasons that are explained in section 6 of this report. This will not happen. It is safe to predict that a comprehensive, global and universal system of QoS-capable interconnection will not happen without some kind of help.

It might nonetheless be possible to get inter-provider QoS to deploy. Anything that can reduce the associated transaction costs will tend to increase the likelihood of getting a decent deployment. Some initiatives that might possibly reduce transaction costs include:

- Experiments and pilot projects among pairs or small groups of cooperating service providers.
- Once the problem is better understood, model agreements for inter-provider interconnection including QoS support.[75]
- The continued enhancement of commercial monitoring and measurement tools that could serve as standardized building blocks for service provider operational support systems (OSS).
- Possible emergence of organizations that could gain acceptance as trusted third parties to capture statistics and/or to mediate billing and accounting disputes.

In addition, it is worth noting that the service providers are unable to require that the entire world implement QoS. Most providers will find that the majority of their traffic is exchanged with a limited number of "trading partners in bits", perhaps a dozen or two. Any realistic provider deployment plan will have to simply accept that some providers will offer QoS-capable interconnection, while others will not.

4 MARKET POWER AND NGN INTERCONNECTION

At the regulatory and policy level, interconnection has always been closely associated with questions of market power. It has been a general article of faith that governments must be prepared to intervene to address such abuses of market power as might exist.

Telecommunications networks were initially presumed to be natural monopolies, industries where initial costs were so high as to preclude competition between two providers in a single geographic area. In most countries, the government itself provided these services, usually through a Post, Telephone and Telegraphy (PTT) authority. In a few, notably including the United States and Canada, equivalent services were historically provided by highly regulated firms that were *de facto* monopolies with significant *de jure* privileges and protection.

With liberalization, services that were previously provided by the government have been privatized, and competitors have been encouraged to enter these markets. In most cases, the established incumbents have resisted competitive entry, either by price-based or by non-price-based discrimination.[76] This behavior is conditioned and shaped by legal and regulatory institutions in each country, but similar underlying economic factors tend to encourage similar incumbent behaviors in all countries.[77]

Once competition is established and effective, it is generally accepted that regulation should be withdrawn. At that point, market forces will channel service provider behavior more effectively than any regulator could hope to.

At the same time, it is important that regulation not be withdrawn *before* competition is effective. Reform-minded New Zealand attempted for many years to operate without a conventional sector-specific regulator. In 2001, they gave it up as a bad job and implemented lightweight institutions approximating the function of a sector specific regulator. Interminable interconnection disputes were the primary reason.[78]

4.1 Sources of market power

Market power most often arises as a result of control of some asset that represents a competitive bottleneck, and that cannot easily be replicated by competitors. In telephony, the primary concern has usually been with “last mile” facilities, which are discussed in the next sub-section. There are other potential bottlenecks that might manifest themselves in specific circumstances, or perhaps more generally in the future – we consider those as well in the subsequent sections.

4.1.1 Last mile considerations

Wired access to the customer premises (e.g. to the consumer’s residence) tends in to be a durable competitive bottleneck throughout the world, but more so in some countries, and in portions of some countries, than in others.

The emergence of NGN access networks may mitigate these concerns, but it is unlikely to eliminate them for the foreseeable future.

In some developed countries, cable television service is sufficiently widespread, and is sufficiently ubiquitously upgraded to carry data and/or telephony, to significantly mitigate the market power of the wired telephony incumbent. Mobile services may also serve as a counterbalance against the market power of the incumbent, including to an increasing degree wireless broadband services. Satellite must also be considered, but it tends to play less of a role for reasons of cost and scalability. Emerging technologies, including broadband over powerline, may play a significant role in the future.

Nonetheless, last mile bottlenecks are likely to be significant for many years to come, and at least portions of most countries are likely to lack effective competition on the last mile. Wherever last mile competitive bottlenecks exist, established operators are likely to find it profitable to restrict or prevent interconnection. Governments and regulators will need to remain alert to this possibility, and must be prepared to intervene if necessary.

4.1.2 Network externality considerations

Last mile bottlenecks tend to be the most commonly noted concern as regards competitive bottlenecks, but they are not the only possible concern.

A body of economic theory argues that, in markets characterized by strong network externality effects, firms with a strong market share of customers will be motivated to have less-than-perfect interoperability and less-than-perfect interconnection.[79]

These concerns have occasionally been relevant to policy in significant ways. They played a large role in the evaluation of the WorldCom-MCI merger and the attempted WorldCom-Sprint merger.[80]

Economic theory does not provide any clear indication as to how large a market share is needed for these effects to motivate action, i.e. to be profitable. At the same time, there is good reason to believe that the world is generally well below that threshold – Internet interconnection today is nearly perfect worldwide, and interconnection disputes are rare.[81]

4.2 Addressing market power

Different countries will have developed different methodologies for addressing market power as it relates to interconnection. In the view of the author, the approach that the European Union adopted in 2003 reflects a particularly forward-looking way to deal with migrations such as that to the NGN.

Under the European regulatory framework for electronic communications, regulators (1) clearly identify a set of relevant markets that could be of interest; (2) determine, using tools borrowed from competition law and economics, whether any firm or group of firms has Significant Market Power (SMP) on such a market; (3) applies a minimally adequate set of *ex ante* (in advance) remedies only to the firm or firms that possess SMP; and (4) removes any corresponding obligations that might have previously existed from firms that do not possess SMP. The framework is technologically neutral – whether a service is delivered using a traditional network or an IP-based NGN is irrelevant. A relevant market is determined based on the service or services delivered to the user, and considering the degree of substitutability for other services, consistent with competition law.

Properly implemented, a regulatory framework of this type enables a regulator to address such market power as may still exist in an NGN world, and also provides a natural and organic method for withdrawing regulation when it is no longer needed.[82]

4.3 Remedies for market power, or a "regulatory holiday"?

In Europe and in North America, a key question has emerged: What is the most appropriate role for government in ensuring that necessary investments are made in new network infrastructure? The debate has largely focused on broadband Internet access, which can be viewed as the access portion of the NGN, but similar issues can be raised about the NGN core.

In a perceptive essay[83], Nicholas Garnham observed that regulatory policy is confused to the extent that it tries to follow multiple economic theories at once, without a way to prioritize or to choose among different and mutually contradictory implications. One of these models is the classical view of competition law and economics, which argues that governments must address such market power as may exist. Another is the Hayekian view, which argues that government must refrain from favoring one solution over another, in order to enable the best to survive – a sort of Darwinian economics. A third is the view of Schumpeter, which argues that progress comes from "creative destruction", and that supracompetitive profits are necessary in order to motivate investment.

The Schumpeterian view is sometimes invoked in support of radical deregulation. The competition law view implies instead that, in problematic markets characterized by non-replicable assets, procompetitive regulation may be needed until effective competition has emerged.

Justus Haucap has characterized this tension of objectives as reflecting a confusion of *deregulation* with *liberalization* – both are much praised, sometimes in the same breath, but they are not the same thing.

Liberalization is a matter of *enabling market entry*, which in some cases implies to need to impose or maintain regulation, not necessarily to eliminate it.[84]

4.3.1 Incentives for providers to deploy

In North America, we have seen the rapid withdrawal of regulation. In Europe, the debate has been expressed in terms of the need for a *regulatory holiday* – a deferral or forbearance from regulation for some period of time in order to spur investment. On both continents, there is support in the law for the sensible notion that regulation should not prematurely be imposed on nascent or emerging services. What is not so clear, unfortunately, is the proper balance between the conflicting Schumpeterian and competition law objectives. Beyond that, what exactly is an emerging service? How long should regulation be deferred? When can an emerging service be said to have emerged?

This debate is likely to be with us for some years to come. Both sides will have adherents, and those adherents are likely to be well funded. It may be some years before the effects can be seen to clearly favor one approach or another.

My personal view is that, in markets that are well established, and where one or more market participants continue to have durable and significant market power, that premature withdrawal of procompetitive regulation is likely to do much more harm than good. Deregulation under those conditions might possibly spur investment by the incumbent operator in the near term, but it will also depress investment by competitive operators. Over time, it seems to me that it is likely to lead to less competition, less innovation and less investment than an effectively regulated system.

4.3.2 Return on Investment (ROI) under conditions of risk

Whatever one's views about deregulation of markets that are not yet competitive, it is clearly appropriate for service providers to make a reasonable return on reasonable investments. For a firm that is subject to regulation, this generally implies a need to compute the *Return on Investment (ROI*) that will be considered to be acceptable for regulatory purposes. Greater risks – as might be expected in connection with migration to the NGN – should be associated with greater expected returns.

Regulators typically determine an appropriate ROI by computing an appropriate *Weighted Average Cost of Capital (WACC*) for the firm. The Weighted Average Cost of Capital (WACC) reflects the cost of equity, the cost of debt, and the company's gearing (a measure of the company's ratio between debt and equity).

The Capital Asset Pricing Mechanism ("CAPM") is a widely used and theoretically well grounded methodology for reflecting risk and its impact on the returns that shareholders should expect. In CAPM, the cost of equity capital is rolled up from three components: (1) the risk free rate; (2) the expected market equity risk premium; and (3) the value of beta for the company in question. The *Risk Free Rate (RFR*) is simply the return that an investor would expect on a risk free investment. The *Equity Risk Premium (ERP*) is a stock-market factor, rather than being company specific, that reflects the degree to which investors expect a higher return for putting money into equity instruments (stocks) than into risk free investments. The beta is a relative measure of the risk that is relevant to the specific firm.

Ofcom, the UK regulator, recently conducted a detailed analysis of the appropriate WACC for British Telecom (BT).[85] Their consultation document provides a very lucid overview of the determination of a WACC for an incumbent provider that is on the verge of a rapid migration to an NGN. They chose to *disaggregate* BT's beta – instead of using a single beta for all of BT, they associated a somewhat lower beta with BT's relatively low risk local loop activities, and a somewhat higher beta with the rest of BT's activities. These different betas then led Ofcom to compute two different WACCs and thus to permit different levels of ROI for different parts of BT.

Ofcom considered various options, but they did not finally resolve the ROI that might be appropriate when BT migrates to an NGN (which BT intends to do on a very accelerated schedule. Ofcom has indicated that BT's risk might be slightly higher for next generation core networks, and significantly higher for next generation access networks, than for BT's current network. Ofcom might address this through further refinements to BT's beta; alternatively, they have raised the possibility of addressing these different levels of risk through a modeling mechanism known as Real Options[86].

4.4 The "network neutrality" debate

A debate has raged in the United States over the past several years over the degree to which providers of broadband Internet access service should be obliged to provide nondiscriminatory access to all content[87] available on the Internet, using any equipment and any application and any protocol that does not harm the network.

In essence, there is increasing concern that new forms of market power might emerge and might be exploited by broadband providers. The concern is exacerbated by the movement of phone companies to also provide video programming, thus offering a vertically integrated service that competes with cable television.

A number of very different concerns have been raised under the banner of network neutrality, mostly in connection with local telephone incumbents or cable TV operators that are vertically integrated with an Internet Service Providers (ISP):

- The possibility that an integrated ISP might offer better performance to some Internet sites than to others;
- The possibility that an integrated ISP might assess a surcharge where a customer wants better-than-standard performance to certain Internet sites;
- The fear that the integrated ISP might permit access only to affiliated sites, and block access to unaffiliated sites;
- The fear that the integrated ISP might assess surcharges for the use of certain applications, or of certain devices;
- The fear that the integrated ISP might disallow outright the use of certain applications, or of certain devices, especially where those applications or devices compete with services that the integrated ISP offers and for which it charges; and
- The fear that the integrated ISP might erect "tollgates" in order to collect unwarranted charges from unaffiliated content providers who need to reach the integrated ISP's customers.[88]

The perceptive reader will have already observed that a number of these concerns (but not all) relate to conduct that, in the absence of market power, would clearly tend to enhance consumer welfare. In a fully competitive market, demanding a surcharge for better performance or for the ability to use highly valued applications would be unobjectionable. With effective competition, the potential for abuse – for example, in the form of assessing chargest that exceed cost to an unreasonable degree – would tend to be contained by the likelihood that competitors would find it profitable to steal customers by offering equivalent services at prices that were less elevated, or under terms and conditions that were less onerous.

As an example, some net neutrality advocates have complained because their provider would offer static (i.e. permanent) IP addresses only in connection with higher-priced services. They complained that they were effectively being prevented from running web servers and other services. In an economic sense, however, this "blockage" is not necessarily problematic. Running a web server will, on the average, result in more traffic for the provider's network, which will in turn tend to result in increased cost to the provider. Aside from that, it represents increased utility to the consumer, and thus an increased surplus and an increased willingness to pay on the part of the consumer. In economics, one of the key properties associated with a service that can be offered for sale is *excludability* – the ability to prevent its use by those who have not paid for it. In this sense, providing static IP addresses only in connection with a higher priced service would, in a competitive marketplace, be viewed as entirely normal and appropriate.

It is also worth noting that there are a great many legitimate reasons to block access to specific Internet addresses – most notably, concerns about security or SPAM. Beyond this, no Internet provider is able to guarantee access to all Internet addresses at all times.

All of this suggests, first, that there is enormous confusion and ambiguity as to what conduct is truly objectionable, and second, that it would be exceptionally difficult to craft a meaningful and enforceable *ex ante* rule to prevent abuse.

4.4.1 Developments in the U.S.

On March 3, 2005, the FCC announced that it had reached a consent decree with Madison River, a small local telecommunications incumbent.[89] Madison River agreed to make a payment, in effect a fine, in recognition that it had blocked access to VoIP services offered by Vonage.

The FCC has not published supporting details,[90] but one might reasonably infer (1) that Madison River customers had little or no ability to choose another broadband provider, and (2) that Madison River chose to block Vonage in order to prevent competition with its own conventional PSTN voice services. If these conjectures are true, then Madison River's conduct was indeed problematic – its actions could be viewed as a leveraging of last mile market power into an otherwise competitive market.

The net result of the FCC's actions, however, must be said to be very confused. The action was, in a sense, probably appropriate, but it left no clear ground rules going forward. Normally, a firm can be fined for willfully violating an FCC rule; however, that implies that there was a rule to violate, and that the company knew or could reasonably infer the rule. The FCC has published no rule, and it is difficult to see how any company could reasonably infer what conduct is permitted and what conduct prohibited today.

Meanwhile, the issue continues to churn in the United States. In recent days, a number of senior telephone company and cable TV executives have spoken of the need to charge content providers such as Yahoo and Google (who are not necessarily customers of th integrated ISP in question) for their use of the ISP's network to reach the integrated ISP's customers. This is not a new idea – it was tried in the past, with no success. In a competitive market, the content providers will simply refuse to pay. An open question is whether recent changes in the U.S. broadband and Internet marketplace, in terms of consolidation and of the collapse of the wholesale market for broadband services,[91] have now made this a profitable strategy.

4.4.2 Policy implications

My view is that there has been very little real abuse of this type to date, and moreover that much of the abuse that has been alleged should not be viewed as problematic. At the same time, there is good reason to believe that problematic behaviors would be both feasible and profitable in the context of a sufficiently concentrated marketplace for broadband Internet access, especially as providers become increasingly vertically integrated.

If these behaviors were to become solidly entrenched, it would be difficult if not impossible to prevent them by means of *ex ante* rules. It is simply too difficult to distinguish between appropriate and inappropriate behavior.

What this strongly suggests is that most countries would be well advised to ensure that they maintain robust competition for broadband Internet services. Competition must be the first, and most critical, line of defense. It is worth noting that the competition need not be facilities-based – service-based competition could be perfectly adequate, as long as the underlying facilities provider cannot constrain the competitive provider's connectivity.

A second implication is, in countries where competition law provides an *ex post* complement to sector-specific regulation, that isolated abuses of this type might be most appropriately addressed *ex post* as violations of competition law, rather than by *ex ante* regulation. My belief is that the truly problematic abuses generally represent inappropriate exploitation of market power.

5 UNIVERSAL SERVICE AND NGN INTERCONNECTION

Charges associated with interconnection are often used as a means of financing universal service – the availability of basic electronic communications to all, at affordable prices. Section 5.1 explains the rationale, in terms of network externalities, economic distortions, and consumer welfare. Section 5.2 explains the use of implicit interconnection-based subsidies within a developing country, while Section 5.3 explores subsidization mechanisms among independent nations. Section 5.4 expands on the implications for policy.

5.1 Network externalities, economic distortions, and consumer welfare

In section 3.x, we explained that markets characterized by network externalities may have a tendency to reach stable equilibrium at levels of service adoption that are much lower than those that are societally optimal. Most countries have felt that voice telephone service was so important that the government should subsidize the service where necessary in order to ensure that the service is available to all, and even to those of limited means. In some cases, this has meant a commitment to universal access (e.g. availability in a nearby school, library or post office) rather than in the home.

Different countries generate these subsidies in different ways. Most economists would argue that it is best to take the funds from general revenues (i.e. overall taxation), because doing so ensures that the cost is spread as widely and as equitably as possible, and thus minimizes economic distortions; however, this is very rarely done in practice.

Some countries simply expect the incumbent local carrier to provide universal service, and to someone extract enough profit from other customers to cover the cost. Still others provide a specific universal service fund, with all providers of electronic communication services contributing.

The relevance of this discussion to interconnection arrangements is that intercarrier compensation is often used as an alternative, implicit means of generating the necessary subsidies.

5.2 Intercarrier compensation as a funding mechanism for ICT development

Domestically, access charges can provide a funding vehicle in the form of implicit subsidies. Network costs will tend to be greater in those areas that pose universal service challenges due to low teledensity or unfavorable geography. Some countries find it convenient to set access charges to higher levels in those areas in order to generate a net influx of money.

The World Bank has generally been supportive of the use of access charges as means of subsidizing telecoms deployment to rural or remote areas of developing countries.

At the same time, this technique is by no means limited to developing countries. It continues to generate implicit universal subsidies in a number of developed countries, including the United States. The U.S. has attempted to phase out these implicit subsidies for years, but they persist.

A number of concerns must be raised in connection with these subsidies. They represent an economic distortion. They are subtle, and not likely to be understood by the public – there can thus be a notable lack of transparency. And they can easily turn into “slush funds”.

5.3 Traffic imbalance – the “Robin Hood” effect

In section 2 of this report, we explained that traditional PSTN intercarrier compensation in most countries is paid according to the Calling Party’s Network Pays (CPNP) principle. It turns out that inhabitants of developed countries tend to place far more calls to inhabitants of developing countries than *vice versa*; consequently, these international termination fees (technically referred to as *settlement fees*) generate a net transfer of money from developed countries to developing countries.

This mechanism has the rather strange property of transferring money from richer countries to poorer ones. As such, one could draw a certain parallel to the mythical English folk hero Robin Hood, who robbed from the rich in order to give to the poor. The system functions as an inadvertent form of foreign aid.

Not surprisingly, developing countries have generally wanted to keep per-minute wholesale termination fees[92] at very high levels, well in excess of real cost, in order to maximize the transfer of funds. Equally unsurprisingly, a number of developed countries, most notably the United States, have wanted to drive these payments down to levels approximating real termination costs.

In one recent incident, the government of Jamaica imposed a levy on international call termination payments, in order to explicitly generate subsidies to fund universal service.[93] The U.S. FCC complained, saying that “… universal service obligations must be administered in a transparent, non-discriminatory and competitively neutral manner, and that hidden subsidies in settlement rates and subsidies borne disproportionately by one service, in the case of settlement rates, by consumers from net payer countries, are not consistent with these principles and cannot be sustained in a competitive global market."[94]

5.4 Policy implications

The migration from today's world of the PSTN to tomorrow's world of the IP-based NGN probably implies that all of these implicit subsidy mechanisms will gradually either be explicitly phased out, or else will become irrelevant over time.

These termination payments are assuredly not an ideal subsidy mechanism; nonetheless, the fact remains that they have transferred funds to developing countries, and that portions of those funds may have served to fund telecoms development projects to remote or rural areas. The funding vehicle is likely to go away, but the development needs that it addressed, however imperfectly, will remain.

6 BILLING AND ACCOUNTING IN AN IP-BASED WORLD

Up to this point, we have primarily considered possible intercarrier compensation arrangements from an economic perspective. These arrangements interact with the underlying IP technology in complicated ways, and have business implications that are perhaps unobvious. In this section, we explore some of the interactions between technology and economics.

6.1 Protocol layering, services, and the underlying network

In an IP-based environment, applications such as Voice over IP (VoIP) operate over an IP-based core network. Protocols are layered in the interest of simplifying the network, and facilitating its evolution over time. These properties have profound implications, not only for usage accounting and billing, but also for the structure of the industry.

Historically, it was generally the case that a single organization would provide both the public telephony *service* and the *network* used to deliver that service. In the world of the IP-based NGN, the network provider will still in most cases still be a service provider, but it will not necessarily be the *only* service provider. Vonage, Skype and SIPgate are examples of competitive firms that provide services without operating a network of their own. For the foreseeable future, integrated and independent service providers are likely to coexist, and to compete for the same end-users customers. Moreover, this competition between integrated and independent service providers is a useful thing, that should be preserved – it tends to enhance consumer welfare.

This separation of function has profound implications for both the network provider and the service provider.

In theory, the network provider in an IP-based world does not know or care about the nature of the application traffic that it is carrying – and in this context, voice is just another application. The network is aware of the Quality of Service that the application has requested for any particular packet, but it should not concern itself with the application itself.

Conversely, the application provider – for example, the independent VoIP provider – will have little or no visibility into the networks that it is traversing. In fact, the application will not necessarily be able to predict which networks its traffic will traverse, and in general the application should not care. The networks collectively provide a path for the application's data traffic, but little more. The application can request a particular Quality of Service for its traffic, but without absolute certainty that its request will be honored.

This lack of awareness has in general proven to be a valuable quality, but it has implications. The independent application provider cannot guarantee the quality of transmission, because it does not own the underlying networks and may not know or care which networks are involved.

The application service space, for example for VoIP, will tend to be a highly competitive market segment unless regulation or anticompetitive actions on the part of network operators (see the discussion on Network Neutrality later in the section) dictate otherwise. The competitiveness of the segment will tend to restrict prices to competitive levels, generally reflecting marginal cost plus a reasonable return on investment. This same competition will tend to constrain the price that the network operator can charge for its integrated service.

All indications are that the marginal cost of the VoIP-based telephony service, independent of the underlying network, is very low.[95] If independent VoIP service providers indeed maintain a competitive market for the service, then the low marginal cost should lead to a low marginal consumer price for the service.

At the same time, the network operator may have (absent regulation) some degree of market power associated with last mile broadband access. To the degree that this is so, the network operator could be said to have market power on one market segment (network access, especially last mile access) that is vertically related to another market segment that is competitive. Under those circumstances, the network operator is likely to exploit its market power, and may try to extend it to the otherwise competitive segment. The simplest and most likely strategy is for the network operator to take a high mark-up (a monopoly profit if it is the only network operator) on the last mile network access, while pricing the voice application at competitive levels.

For this reason, many countries will find it necessary to maintain regulation that seeks to address durable bottlenecks associated with last mile access, to the extent that effective competition has not yet emerged for the last mile. Countries will see these needs through the lens of their own experience and their own institutions, but many or most will find it necessary to retain regulatory measures, or to institute them if they do not exist, in order to enable competitive entry and to sustain it over time, and to limit the exploitation of market power where competition is not yet effective.

6.2 Point-to-point versus end-to-end measurement

The technology and economics of these systems interact in complicated ways.

The underlying network economics strongly influence the nature of the things that operators and service providers will want to bill for; however, those bills will have to be justified and reconciled based on some kind of accounting data. Billing needs largely determine accounting system needs.

Conversely, not all of the data that might be desired can be acquired at reasonable cost, so the capabilities that can reasonably be achieved by accounting systems necessarily reflect back and influence what metrics could potentially be used for billing.

In the wired PSTN, the points of origination and termination are generally known or knowable when the call is initiated. Once the call is initiated, these points remain stable for the duration of the call. The traffic during the call is not relevant to the bill. Typically, the only accounting datum needed after the call has been originated is the time at which the time at which it ends.

In the Internet, some things are known at the level of the *application* or *service*, while very different things are known at the level of the *network*. For VoIP, a server that implements a protocol like SIP will know the time at which a session is initiated, and may know that time at which it ends, but will know next to nothing about the network resources consumed in the interim. The *topological* location (the logical location within the network) of the originating and terminating end points will be known, but not necessarily the *geographical* location.[96]

Beyond this, an IP-based network will be dealing with a far broader array of applications than just traditional voice. The notion that the *call originator* should be viewed as the *cost causer* breaks down in the general case. In the general case, there is no obvious "right answer" to the question of how to allocate costs among end-users.

The underlying *network* knows very different things. In an IP-based environment, each IP datagram is independently addressed, and could in principle be independently routed (although routing in practice is much more stable than this implies). Relatively simple applications can generate a very large number of IP datagrams. For accounting purposes, it is necessary to summarize this data – otherwise, the accounting systems will be deluged with unmanageable data volumes.

For analogous reasons, it is trivial to measure the traffic over a given point-to-point data transmission link, but expensive and cumbersome to develop an overall traffic matrix based on end-to-end traffic destinations.

For all of these reasons, billing and accounting arrangements in the Internet have historically tended to reflect huge simplifying assumptions. For individual consumes and for enterprise customers, billing has most often been on a flat rate basis, as a function of the maximum capacity of the access link from the service provider to the customer (i.e. the price is based on the size of the "pipe", which sets an upper limit on the amount of traffic that the provider must carry).

At an enterprise level, prices have sometimes reflected the total traffic carried over the pipe, most often based on some percentile of data transmission rates (for example, a 95th percentile of rates sampled at 15 minute intervals, which will correspond roughly to average traffic for the busiest hour of the day).

It is important to note what is *not* charged for. Network operators do not assess usage-based charges for things that they cannot measure (at reasonable cost). Retail prices do not generally reflect either the distance that IP-based traffic is carried, or the degree to which international boundaries are crossed. It is simply too difficult and too expensive to measure these things. Wholesale arrangements between providers might take account of distance to some extent – the providers know the circuits between them, and can measure the point-to-point traffic over those circuits.

6.3 Reconciliation of statistics

To the extent that billing reflects usage, occasional issues and disagreements are inevitable. It is important that providers be able to reconcile their usage statistics, and that they be able to reach agreement at reasonable cost.

At the retail level, providers often choose to avoid this issue entirely by avoiding usage-based prices. At the wholesale level, the use of Bill and Keep peering arrangements also serves to reduce if not eliminate the need to reconcile statistics.

Where two providers charge one another based on traffic sent in both directions, reconciliation will be necessary. One might well imagine that, where provider A measures the traffic over a particular transmission link to provider B, that that measurement should correspond exactly to B's measurement of traffic from A to B over the same transmission link. My experience during my time in industry suggests, unfortunately, that disputes will occasionally occur, even where both parties are (most likely) acting in good faith, and even where it would seem that both parties should be measuring the same thing.

There are steps that can be taken to reduce, but not prevent, misunderstandings. Coordinating reporting start times and intervals can help. This is particularly important if the usage charges between providers depend on a percentile measure of traffic – the mean of traffic is independent of sampling interval, but the standard deviation is not. Sampling a given stream at more frequent intervals will lead to a "lumpier" distribution – a fundamental consequence of the Central Limit Theorem. If two organizations want to reach the same conclusions about a percentile, they should sample with identical frequency.

An approach that has sometimes been used – for example, in the U.S. mobile industry at one point – is to have a trusted intermediary collect and analyze the statistics. In general, the intermediary cannot itself be a competitor in the same market – otherwise, it will not be trusted.

6.4 Accounting for Quality of Service[97]

If two providers want to compensate one another for carrying their respective delay-sensitive traffic at a preferred Quality of Service, each will want to verify that the other has in fact done what it committed to do.

In the case of QoS, this would seem to imply measurements of (1) the amount of traffic of each class of service exchanged in each direction between the providers; and (2) metrics of the quality of service actually provided. Measuring the volume of traffic by class is, once again, trivial – it is no harder than measuring the overall traffic for the same transmission link. Measuring the QoS is much more complex, both at a technical level and at a business level.

For QoS, commitments between providers would presumably be primarily in terms of the mean and variance of delay. One can measure delay with primitive tools such as PING[98], or with more sophisticated tools such as IPPM probes.[99] One could imagine a pair of providers who mutually agree to instrument their networks to support one or more of these measurement tools, and to mutually measure delay between their respective networks. One might imagine that this should be easy – one would need to agree where the probe points should be physically situated, and what measurement metrics should be employed, and one might imagine that nothing more should be needed. The reality is much more complex.

First, it is important to remember that this measurement activity implies a degree of cooperation between network operators who are direct competitors for the same end-user customers. Each operator will be sensitive about revealing the internal performance characteristics of its networks to a competitor. Neither would want the other to reveal any limitations in its network to prospective customers.

Second, there might be concerns that the measurement servers – operated within one's own network, for the benefit of a competitor – might turn into an operational nightmare, or perhaps a security exposure, within the perimeter of one's own network.

Again, there might possibly be scope for a trusted and independent third party to perform this function.

6.5 Gaming the system

If the arrangements between providers were such as to make it attractive to carry delay-sensitive traffic, then it is safe to predict that some providers will attempt, absent countermeasures, to benefit from the arrangements. Whether this should be viewed as fraud, as arbitrage, or simply as creative entrepreneurship might depend on the specific circumstances, and might be difficult to judge in practice.

For example, a network operator might discount its retail connectivity prices to end-user enterprises that operate call centers, on the theory that the resulting traffic would enable it to capture more revenue from other operators for carrying high-QoS traffic. This would seem to be a legitimate business option.

On the other hand, one could imagine an operator creating, or causing to be created, a software robot that would generate a great deal of otherwise unnecessary traffic that the operator would then have to be paid to deliver. This would seem to be a matter of arbitrage or worse, with no redeeming value.

In practice, distinguishing between appropriate and inappropriate arrangements is likely to be difficult. The actual forms that abuse might take cannot be predicted with confidence.

7 A Hypothetical Scenario: Interconnection in an NGN world

In this section, we consider possible consequences of the migration to an IP-based NGN. It is a thought experiment that seeks to shed light on possible developments.

We develop a scenario, premised on the assumption that the primary incumbent in a country that operates within the regulatory framework of the European Union migrates to an IP-based NGN core.

The country is assumed, on the eve of migration, to have:

(1) an incumbent wired and wireless operator that had previously been the country's PTT, and that still has substantial market share and market power;

(2) various wired and wireless competitve operators;

(3) various independent providers of broadband Internet services, some facilities-based, some providing service competition based on procompetitive regulation (LLU, bitstream, and shared access);

(4) several independent providers of VoIP; and

(5) a number of local providers of Internet content, both web and video.

Our focus here is on IP-based NGN core migration. The characteristics of NGN access migration are, for these purposes, assumed to be possibly different in scale but similar in concept to the broadband deployment that we see today.

I have attempted to sketch a number of plausible scenarios, but I must emphasize at the outset that this is a highly speculative and perhaps controversial business. As the American baseball coach Yogi Berra once said, "It's hard to make predictions, especially about the future."

7.1 The scenario

During an extended transitional phase, the historic incumbent (BigCo for purposes of this discussion) operates traditional PSTN-based voiced services, traditional broadband and dial-up Internet access, and new integrated IP-based NGN capabilities. The NGN-based capabilities are first offered opportunistically in those areas where demand is expected to be highest and most concentrated, or in areas that required significant upgrades independent of the migration to NGN.

In the longer term, the migration to NGN will enable BigCo to achieve not only faster time-to-market for new services, but also cost savings through integration. In the near term, however, unit costs may tend to be stable or possibly to *increase*, for two reasons. First, it is unlikely to be cost-effective to decommission much of the current network until the migration is quite far advanced; and second, the need to operate two kinds of infrastructure in parallel during the transition implies increased operational expense for engineering, training, spare parts, support and operations staff, and the maintenance of software operational support systems.

Assuming a competitive retail market, BigCo is unlikely to increase prices in response to any short term increase in unit costs. They will not want to lose hard-to-replace customers to competitors. A more likely scenario is that they will hold prices steay or reduce them slightly, effectively subsidizing current customers by borrowing from anticipated future savings.

BigCo’s traditional competitors will respond to perceived competitive pressure by initiating their own migration to NGN core networks, if they have not already done so. This will be prompted in part by the need to achieve economies of scale and scope closer to those of BigCo, and partly by the fear that they will otherwise be unable to compete when BigCo is eventually permitted to withdraw regulatorily mandated traditional PSTN interconnection in favor of NGN interconnection.

IP-based competitors will not perceive the need to make radical changes to their operations – they are, for the most part, already there. They will perceive a need to anticipate forthcoming IP-based NGN interconnect offerings.

As the transition phase comes to a close, BigCo will phase out traditional services on a large scale. From this point forward, the traditional services and traditional models of interconnect become less relevant.

7.2 Regulatory implications for last mile access

During the transition phase, existing regulatory oblgations for access to last mile facilities, both for traditional PSTN-based competitors and for broadband providers, will likely need to be maintained. In the near term, the last mile will continue to represent a durable competitive bottleneck in most (but not all) regions of most countries. In the near term, neither the migration to an NGN core nor the incumbent’s deployment of NGN access will obviate the need for competitive access. In other words, BigCO will most likely continue to possess whatever last mile market power it had prior to the migration to NGN. In the European context, this implies the continuation of some combination of local loop unbundling (LLU), shared access, bitstream access, and resale.

For countries, or regions of countries, where three or more effective facilities-based alternative broadband options are available, and to the extent that competition appears to be effective and sustainable, it may be appropriate to eliminate or phase out these last mile obligations.

When migration is well advanced, it is possible that broadband competition will be the only meaningful last mile competition that is meaningful. There may be no further need to enable resale or LLU as an enabler for PSTN-based competition.

7.3 Regulatory implications for interconnection

During the transition phase, BigCo will still be obliged to maintain traditional PSTN interconnection capabilities. Assuming that it is possible for competitors to reach BigCo’s NGN-based end-user customers through traditional interconnection, there will not necessarily be a regulatory obligation to provide new NGN-based interconnection capabilities.

BigCo will offer IP-based interconnection at some point during the transition phase. As the transition phase draws to a close, they will want to withdraw traditional interconnection. To the extent that they still possess market power, they will almost certainly be under regulatory obligations to provide NGN interconnection at cost-based prices. To the extent that the NGN implies lower forward-looking unit costs, the cost-based interconnection prices will be lower than those that pertain today.

In Europe today, all or nearly all operators that provide publicly available telephone service (PATS) are subject to regulatory obligations to interconnect, because all – even small operators, as we have seen in section 2 of this report – have significant market power in regard to the termination of telephone calls.

7.4 Peering versus transit

As we have seen, in the world of the Internet, the great majority of interconnection take the form either of peering or of transit. In our hypothetical scenario, will market participants prefer peering, transit, or some other model of interconnection? Recall that peering offers exchange of traffic only between BigCo's customers and those of its peer, but does not provide either with access to third parties. In a typical transit relationship, by contrast, the transit customer can use the transit provider's network to reach destinations anywhere on the Internet.

7.4.1 Peering versus Transit for international interconnection

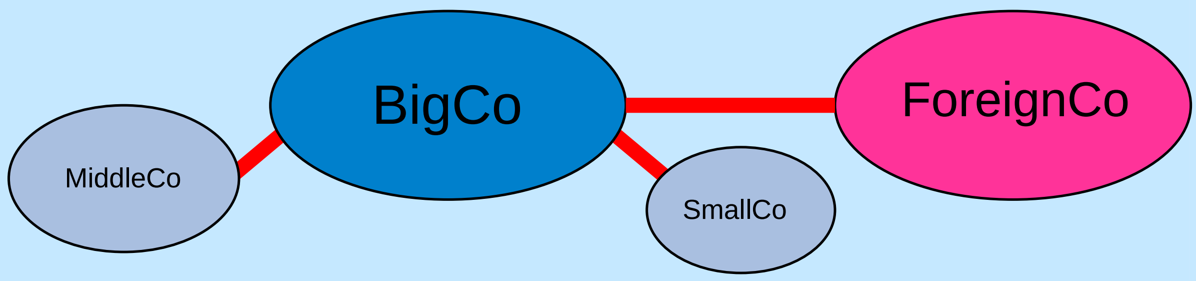
We start by considering BigCo's relationship to similarly situated operators in other countries. Experience to date strongly suggests that these arrangements will tend to be peering relationships. Historically, peering arrangements have usually been on a Bill and Keep basis; however, in an NGN world that supports differentiated QoS, it is possible that BigCo and its peer might agree to one level of charges for conventional best efforts traffic and another, higher level of charges for traffic with preferred QoS. In fact, there could be more than two levels.

On the other hand, BigCo is unlikely to agree to peer with tiny competitive operators, either in other countries or for that matter in BigCo's own country. This implies that tiny, competitive operators will generally need to contract with some transit provider (but not necessarily BigCo).

There is likely to be an extended period of coexistence, where BigCo interconnects with some operators (especially foreign operators) by peering, with others by transit, and with quite a few others by means of traditional PSTN interconnection. Internationally, traditional PSTN interconnection will surely persist.

There is also a matter of transaction costs – each interface migration from a PSTN basis to an NGN/IP basis implies certain real transition costs, as well as transaction costs associated with creating and managing new interconnection agreements. Overnight mass migration cannot be cost-effective. This implies that BigCo will, other things being equal, first seek out IP-based interconnection arrangements with those operators with which the agreements provide it with the greatest benefit, which might tend to be those similarly situated operators with which it exchanges the largest volume of traffic.

Figure 7.1: Hypothetical peering arrangements



The transition costs pose a regulatory challenge as well. To the extent that BigCo unilaterally chooses to massively re-shape its network in the NGN world, possibly withdrawing network interconnection points, what are its obligation to competitive providers with which it has existing arrangements? It seems inappropriate that competitive providers should be involuntarily burdened with new costs that are not of their making; at the same time, BigCo should not be forced to maintain obsolete interconnection points indefinitely. These complicated trade-offs have been a central theme in several Ofcom (UK) public consultations on the migration to the NGN.[100]

Finally, we note that an incentives problem could easily arise that could slow or prevent the migration to next generation international interconnections. The existing arrangements tend to transfer significant sums of money from one operator to another, either because mobile rates are much higher than fixed, or because far more calls are initiated from developed countries to developing ones than *vice versa*. The migration to peering is likely to result either in Bill and Keep or in cost-based arrangements, which would either reduce or eliminate the subsidies. This means that two operators that contemplate a migration from current arrangements to IP-based peering are likely to perceive the change as a zero-sum game – one provider will benefit from the change, and one will suffer. Under those assumptions, the provider that is negatively impacted can reasonably be expected to refuse to make the transition, or, if somehow compelled to upgrade, to delay the transition as long as possible.

7.4.2 Peering versus transit for domestic interconnection within BigCo's country

As previously noted, BigCo is unlikely to be motivated to offer peering arrangements to tiny competitive operators in its own country. It might offer peering arrangements to just a few of its largest domestic competitors.

A difference between this case and the international case is that these competitive operators will be highly motivated to have good connectivity to BigCo's customers. (To the extent that BigCo's customer base is much larger than that of its competitors, it will tend to prefer less-than-perfect interconnection with small competitors. This is a straightforward application of the Katz-Schapiro result discussed in section 2 of this paper.[101])

At that point, small domestic competitors have limited options:

(1) As long as traditional PSTN interconnect is offered, and to the extent that it is sufficient for the competitor's needs, they might stick with PSTN interconnect.

(2) They can purchase transit service from BigCo.

(3) They can purchase transit service from some provider other than BigCo.

My prediction is that many of the small domestic providers would choose to purchase transit service from BigCo (perhaps in addition to service from some other transit provider) as long as BigCo's price is competitive.

As long as the market for wholesale transit services is reasonably competitive (and assuming that BigCo also faces an effectively competitive market for broadband Internet access), this should lead to quite reasonable domestic outcomes. BigCo's wholesale price for transit service will be constrained by competition from third parties. BigCo's competitors need access to BigCo's customers, and will prefer the best connection that they can afford, *but they can reach BigCo's customers perfectly well through a third party transit provider*.

This is an important distinction between the NGN world and the PSTN world. In the IP-based world, indirect interconnection is perfectly reasonable.

To the extent that peering arrangements with domestic competitors either are on a Bill and Keep basis, or that they reflect roughly balanced net payments,[102] and to the extent that underlying facilities are available on a competitive or a nondiscriminatory basis, the competitors' costs to reach BigCo's customers should not greatly exceed those of BigCo itself (except to the extent that BigCo enjoys advantages of scale). Consequently, competition from these domestic competitors should appropriately constrain BigCo's behavior, and prices are likely to be competed down to levels not greatly in excess of marginal cost.

Foreign peers would experience somewhat higher costs in competing for BigCo's domestic end user customers, but only to the extent that their costs are impacted by lacking a local base of operations.[103]

Potential competition from foreign service providers thus provides a second (albeit looser) constraint on BigCo's pricing power. If BigCo were to attempt to price well in excess of cost, these foreign providers might be motivated to establish a presence in BigCo's country so as to compete directly.

To re-cap, this implies that the likely domestic pattern is:

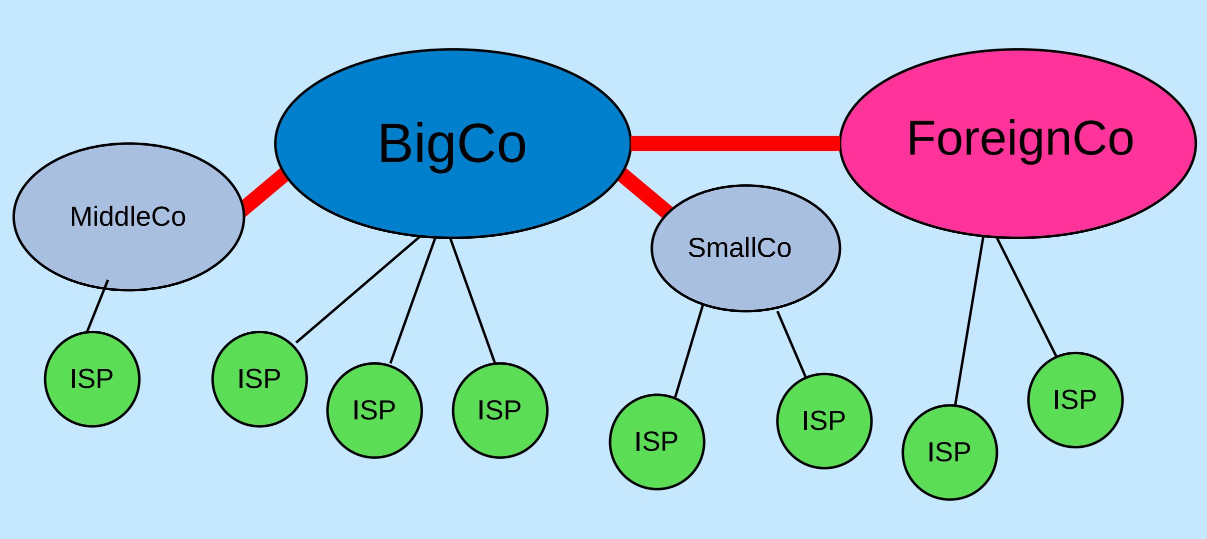
(1) a few of the largest competitors might peer with BigCo;

(2) small competitors will purchase transit from BigCo if they can;

(3) small competitors will supplement or replace BigCo transit with transit from third parties; and

(4) small competitors may choose, as an economic optimization, to peer with one another whenever the traffic that they can exchange reduces their transit costs sufficiently to pay for the cost of any peering circuits and infrastructure.[104]

Figure 7.2: Hypothetical peering and transit arrangements



This returns us to a key question of regulatory policy. It is natural to assume that BigCo's existing PSTN market power as regards interconnection will automatically confer market power as regards interconnection in the NGN world, and that any interconnection remedies therefore need to automatically carry over to NGN interconnection; I would argue, however, that making this presumption today would be greatly premature. For the reasons outlined above, it is entirely possible (given adequate competition or effective regulatory access to necessary underlying facilities such as leased lines, wholesale transit and broadband Internet access) that unregulated IP-based interconnection will lead to a perfectly satisfactory Coasian solution – a solution which would likely be superior to anything that a regulator could craft.

7.5 Network provider versus application servce provider

In the world of the NGN, the terminating monopoly requires some re-thinking. The end-user may get his or her broadband connection from BigCo, or from a competitive broadband Internet access provider. He may get his voice telephony service – assuming that the service continues to look much as it does today – from BigCo, or he may get it from an alternative VoIP service provider. For telephone calls, if anyone possesses a termination monopoly, it is the VoIP service provider, not the provider of the broadband pipe.

Who, then, should collect the termination charge? It is important to remember that termination costs exist to recompense the terminating carrier for the incremental usage-based costs imposed on its network. An independent VoIP provider has no network, and experiences very little incremental usage-based cost.

Recall, too, that the network provider has only limited visibility into the traffic that it is carrying. The network provider could, however, assess a surcharge for packets where the user explicitly requests preferred

Quality of Service; however, if the charge is high, the user will probably prefer services that operate with standard best-efforts QoS (which will, as previously noted, still provide perfectly adequate voice quality in general). The network operator could conceivably attempt to monitor the user's service in order to assess a surcharge for voice traffic (leaving aside for the moment the possible invasion of privacy that this implies), whether associated with preferred QoS or not; however, if the surcharge were large, users might again respond by encrypting their traffic to prevent the network provider from inspecting it. Technology could conceivably close any or all of these holes, but there is no obvious social benefit in doing so. To the contrary, consumer welfare would appear to be maximized by giving consumers as much latitude as possible to do what they want to do, with as few restrictions as possible.

It also bears noting that it costs the network no more to carry a VoIP packet (on a best efforts basis) than it does to carry a WorldWide Web packet, or any other data packet for that matter. Moreover, the marginal usage-based cost per packet is very, very low.

Yet another challenge relates to cost causation. Historically, it has been assumed that party that originates the call is the sole cost causer. This assumption has always been questionable. Going forward, it will be difficult if not impossible to ascribe cost to one or another party to a communication.

7.6 Implications for differentiated Quality of Service

Within individual IP-based networks, differentiated QoS has existed for many years.

If BigCo prices Internet transit competitively, many competitive operators are likely to choose to procure transit service from BigCo. This positions BigCo to offer QoS-capable access to its competitors, not only to BigCo's own customers, but also to the customers of most domestic competitors.

For reasons noted in section 3 of this paper, inter-provider QoS has been slow to deploy in connection with peering interconnection. Paradoxically, offering it in connection with transit service could be less problematic, provided that it is offered at a price that is not disproportionate to the benefits that it provides. In this scenario, the network externalities advantage that BigCo enjoys by virtue of its large customer base positions it to provide QoS capable transit to most or all competitors on the national market.

This is not a model that a regulator will hasten to embrace, since it implies a unique role for the country's historic incumbent provider. Given the limited benefits that differentiated QoS confers, however, it might represent a quite reasonable trade-off. Whatever market power these arrangements confer on BigCo in regard to QoS would appear to be of limited value.

At the same time, these arrangements do not necessarily lead to a global NGN with ubiquitous support for differentiated QoS. Transaction costs are likely to continue to inhibit implementation of diffentiated IP QoS at the level of peering relationships; consequently, differentiated QoS at the international level is likely to have at best a spotty availability for an extended period of time, even in the event that most service providers ultimately migrate to NGN and to IP-based NGN interconnection.

7.7 Policy implications

With all of this in mind, my view is that interconnection arrangements in an NGN world are likely to be most rational and sustainable to the extent that they adhere to a few guiding principles:

(1) Wherever competitive conditions warrant, a Coasian solution reflecting market-based negotiations between the NGN operators is likely to lead to more efficient solutions than a regulatory rate-setting.

(2) National regulatory authorities might therefore be well advised to focus their attention primarily on ensuring adequate competition for wholesale Internet transit services, and for consumer broadband Internet access.

Where a Coasian resolution is not feasible, the following considerations follow from the previous discussion:

(3) The wholesale charge assessed should either be zero (i.e. Bill and Keep), or should be no higher than the forward-looking marginal usage-based cost associated with carrying the incremental traffic.

(4) As a corollary, incremental charges are appropriate only to the extent that they are associated with incremental costs.

(5) Charging should reflect only things that can be measured in a straightforward and documentable way by the party that assesses the charges.

(6) Charges could reasonably consider the volume of traffic exchanged at each level of QoS requested (and delivered), but should otherwise be independent of the nature of the application employed by the user.

ENDNOTES

[1] Charles Dickens, A Christmas Carol.

[2] See http://www.btglobalservices.com/business/global/en/business/business_innovations/issue_02/century_network.html.

[3] See Haucap, J., and Marcus, J.S., “Why Regulate? Lessons from New Zealand”, IEEE Communications Magazine, November 2005, available at: http://www.comsoc.org/ci1/Public/2005/nov/ (click on "Regulatory and Policy").

[4] See Ofcom’s Final statements on the Strategic Review of Telecommunications, and undertakings in lieu of a reference under the Enterprise Act 2002, September 22, 2005.

[5] “ECTA comments on NGN public policy”, November 2005.

[6] Section 7 benefits from recent developments and from regulatory proceedings in the UK, but the scenario is not patterned after the proposed BT evolution, nor after specific developments in any particular country.

[7] Jean-Jacques Laffont, Patrick Rey and Jean Tirole, “Network Competition: I. Overview and Nondiscriminatory Pricing” (1998a), Rand Journal of Economics, 29:1-37; and “Network Competition: II. Price Discrimination” (1998b), Rand Journal of Economics, 29:38-56.

[8] Armstrong, M. “Network Interconnection in Telecommunications.” Economic Journal, Vol. 108 (1998), pp. 545–564.

[9] Jean-Jacques Laffont and Jean Tirole, Competition in Telecommunications, MIT Press, 2000.

[10] I should hasten to add that I myself am not formally trained as an economist.

[11] See Federal Communications Commission (FCC) Office of Strategic Planning and Policy Analysis (OSP) Working Paper 33: Patrick deGraba, “Bill and Keep at the Central Office As the Efficient Interconnection Regime”, December 2000, available at http://www.fcc.gov/osp/workingp.html.

[12] “On the receiver pays principle”, RAND Journal of Economics, 2004.

[13] Andrew Odlyzko, “The evolution of price discrimination in transportation and its implications for the Internet”, Review of Network Economics, vol. 3, no. 3, September 2004, pp. 323-346, available at http://www.rnejournal.com/articles/odlyzko_RNE_sept_2004.pdf.

[14] Cf. FCC, 8th CMRS Competition Report, §94: “AT&T Wireless’s Digital One Rate (“DOR”) plan, introduced in May 1998, is one notable example of an independent pricing action that altered the market and benefited consumers. Today all of the nationwide operators offer some version of DOR pricing plan which customers can purchase a bucket of MOUs to use on a nationwide or nearly nationwide network without incurring roaming or long distance charges.” Several mobile operators offer a variant of this plan where there are no roaming charges as long as the customer is using that operator’s facilities.

[15] These flat rate plans are truly flat rate, whereas the mobile plans are generally two part tariffs. The usage charges of the mobile plans are usually set to very high levels (as much as $0.40 per Minute of Use). They are not so much intended to be used, as to punish consumers who purchase bundles that are too small. The common feature between the mobile plans and the newer truly flat rate plans is a movement away from meaningful usage charges.

[16] For example, Vonage offers unlimited calls to or from the U.S. and Canada for just $24.99 a month. See www.vonage.com.

[17] In the United States, by means of the Enhanced Service Provider (ESP) exemption; in the UK, by means of FRIACO.

[18] This definition is adapted from Laffont and Tirole (2001), page 182.

[19] In the interest of simplicity, we will gloss over the historically important distinction between access charges and reciprocal compensation in the United States. As the industry consolidates (with the disappearance of AT&T and MCI as independent long distance carriers), this distinction is somewhat less relevant than it once was. For a more detailed treatment of arrangements in the U.S., see Marcus, “Call Termination Fees: The U.S. in global perspective”, presented at the 4th ZEW Conference on the Economics of Information and Communication Technologies, Mannheim, Germany, July 2004. Available at:

ftp://ftp.zew.de/pub/zew-docs/div/IKT04/Paper_Marcus_Parallel_Session.pdf.

[20] In 2001, the FCC signaled its intent to migrate to a much broader implementation of Bill and Keep; however, this regulatory policy change has been stalled for years. See FCC, In the Matter of developing a Unified Intercarrier Compensation Regime, CC Docket 01-92, released April 27, 2001.

[21] See Laffont, Rey and Tirole (1998a) and (1998b); Armstrong (1998); Laffont and Tirole (2001), all op. cit. See also Cave et. al. (2004); de Bijl et. al. (2004); and Haucap and Dewenter (2004).

[22] See FCC OSP Working Paper 33: Patrick DeGraba, “Bill and Keep at the Central Office As the Efficient Interconnection Regime”, and Working Paper 34: Jay M. Atkinson, Christopher C. Barnekov, “A Competitively Neutral Approach to Network Interconnection”, both December 2000, both available at http://www.fcc.gov/osp/workingp.html; Stephen C. Littlechild, “Mobile Termination Charges: Calling Party Pays vs Receiving Party Pays”, forthcoming, available at http://www.econ.cam.ac.uk/dae/repec/cam/pdf/cwpe0426.pdf; Robert W. Crandall and J. Gregory Sidak, “Should Regulators Set Rates to Terminate Calls on Mobile Networks?”, Yale Journal on Regulation, 2004; and Marcus (2004), op. cit.

[23] Laffont and Tirole, Competition in Telecommunications (2001), page 186. The italics are theirs. See also Haucap and Dewenter (2005).

[24] There are, of course, numerous exceptions and caveats to this statement. See chapter 5 of Laffont and Tirole (2001).

[25] See Martin Cave, Olivier Bomsel, Gilles Le Blanc, and Karl-Heinz Neumann, How mobile termination charges shape the dynamics of the telecom sector, July 9, 2003; Paul W.J. de Bijl, Gert Brunekreeft, Eric E.C. van Damme, Pierre Larouche, Natalya Shelkoplyas, Valter Sorana, Interconnected networks, December 2004; Littlechild (2006); and Marcus (2004).

[26] See European Commission, 10th Implementation Report (December 2004); and Marcus (2004)

[27] Laffont and Tirole (2001), page 190.

[28] Milgrom et. al. suggest that this is the economically predicted result for Internet backbones. See Paul Milgrom, Bridger Mitchell and Padmanabhan Srinagesh, “Competitive Effects of Internet Peering Policies”, in The Internet Upheaval, Ingo Vogelsang and Benjamin Compaine (eds), Cambridge: MIT Press (2000): 175-195. At: http://www.stanford.edu/~milgrom/publishedarticles/TPRC%201999.internet%20peering.pdf.

[29] To understand the motivation for this, see Laffont and Tirole (2001) pages 201-202.

[30] An operator might choose to ignore a termination fee that constitutes only a small fraction of the total cost of the call. Termination fees set in the absence of regulation often represent the preponderance of the total cost of the call.

[31] In support of this interpretation, it is worth noting that flat rate mobile plans are common in Europe, but generally only for on-net calls and for calls to the fixed network. The calls that are excluded are precisely the calls to off-net mobile subscribers – the calls where termination fees would tend to be high.

[32] See Federal Communications Commission (FCC) Office of Strategic Planning and Policy Analysis (OSP) Working Paper 33: Patrick DeGraba, “Bill and Keep at the Central Office As the Efficient Interconnection Regime”, December 2000, at 95, available at http://www.fcc.gov/osp/workingp.html.

[33] See footnote 69 on page 28.

[34] Mobile termination fees in the European Union are increasingly subject to regulation, but this is still in the process of being phased in. The fees today continue to reflect to a significant degree the previous unregulated arrangements.

[35] See FCC, Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, 10th Report (10th CMRS Competition Report), July 2005, Table 10, based on Glen Campbell et al., Global Wireless Matrix 4Q04, Global Securities Research & Economics Group, Merrill Lynch, Apr. 13, 2005.

[36] Cf. Crandall and Sidak (2004).

[37] The per-minute fees in US “bucket of minutes” plans are probably not exercised much in practice. They are so high, in comparison with the bucket of minutes arrangements, as to serve primarily as a punitive measure to force users to upgrade to a larger bucket. At the same time, this implies that, for a given user over time, consuming more minutes will equate to higher charges.

[38] In analyzing European experience, Cave et. al. (2003) find that only a small portion of the subsidy is returned to the consumer.

[39] One might well expect a corresponding tendency for CPP/CPNP arrangements to slow the ongoing adoption of the services from which the subsidies are generated, that is, fixed services. Eastern European experience might possibly support this view – for example, mobile phone penetration in Hungary is more than 70% and growing, while fixed phone penetration is about 40% and declining. I am not aware of any rigorous analysis on this question.

[40] See, for example, Cave et. al. (2003); Littlechild (2006); and Crandall and Sidak (2004), all op. cit.

[41] The penetration is usually computed by dividing the number of subscriptions by the population. For penetration to exceed 100% implies that some consumers have more than one subscription at the same time. This probably reflects either (1) pre-paid cards that are nominally active but no longer being used, or (2) consumers who find it cost-effective to place on-net calls on more than one network.

[42] Crandall and Sidak (2004), op. cit.

[43] Report of the NRIC V Interoperability Focus Group, “Service Provider Interconnection for Internet Protocol Best Effort Service”, page 7, available at http://www.nric.org/fg/fg4/ISP_Interconnection.doc.

[44] Ibid., pages 4-6. See also Marcus, Designing Wide Area Networks and Internetworks: A Practical Guide, Addison Wesley, 1999, Chapter 14.

[45] The current number is probably far less.

[46] See http://www.apnic.net/services/asn_guide.html.

[47] A very innovative paper by Prof. Gao of the University of Amherst confirms this structure. See Lixin Gao, "On inferring autonomous system relationships in the Internet," in Proc. IEEE Global Internet Symposium, November 2000. The Internet is probably more richly interconnected today than was the case in 2000, but there is no reason to believe that these basic aspects have changed very much.

[48] Paul Milgrom, Bridger Mitchell and Padmanabhan Srinagesh, “Competitive Effects of Internet Peering Policies”, in The Internet Upheaval, Ingo Vogelsang and Benjamin Compaine (eds), Cambridge: MIT Press (2000): 175-195. At: http://www.stanford.edu/~milgrom/publishedarticles/TPRC%201999.internet%20peering.pdf.

[49] See M. Katz and C. Shapiro (1985), “Network externalities, competition, and compatibility”, American Economic Review 75, 424-440.; and J. Farrell and G. Saloner (1985), `Standardization, compatibility and innovation', Rand Journal of Economics 16, 70-83.

[50] Jacques Cremer, Patrick Rey, and Jean Tirole, Connectivity in the Commercial Internet, May 1999.

[51] Milgrom et. al., “Competitive Effects of Internet Peering Policies” (2000).

[52] Private communication, Marius Schwarz, Georgetown University.

[53] Armstrong, M. “Network Interconnection in Telecommunications.” Economic Journal, Vol. 108 (1998), pp. 545–564.

[54] Laffont, J.-J., Rey, P., And Tirole, J. "Network Competition: I. Overview and Nondiscriminatory Pricing." RAND Journal of Economics, Vol. 29 (1998a), pp. 1–37.

[55] Laffont, J.-J., Marcus, J.S., Rey, P., And Tirole, J., "Internet interconnection and the off-net-cost pricing principle", RAND Journal of Economics, Vol. 34, No. 2, Summer 2003, available at http://www.rje.org/abstracts/abstracts/2003/rje.sum03.Laffont.pdf. A shorter version of the paper appeared as "Internet Peering", American Economics Review, Volume 91, Number 2, May 2001.

[56] This is not to suggest that all parties have been satisfied with the results. An ongoing dispute over International Charging Arrangements for Internet Service (ICAIS) has been simmering for some years now.

[57] Ronald H. Coase, "The Federal Communications Commission", Journal of Law and Economics 2 1-40, 1959.

[58] Industry consolidation is another noteworthy contributory factor.

[59] FCC, In the Matter of developing a Unified Intercarrier Compensation Regime, CC Docket 01-92, released April 27, 2001, section 95. See also Marcus, "Call Termination Fees: The U.S. in global perspective", July 2004, available at: ftp://ftp.zew.de/pub/zew-docs/div/IKT04/Paper_Marcus_Parallel_Session.pdf.

[60] Andrew Odlyzko has written a number of insightful papers exploring the historical roots of price discrimination, and the relevance to the Internet. See Andrew Odlyzko, "The evolution of price discrimination in transportation and its implications for the Internet", Review of Network Economics, vol. 3, no. 3, September 2004, pp. 323-346, available at http://www.rnejournal.com/articles/odlyzko_RNE_sept_2004.pdf.

[61] See the classic paper by the Stanford University mathematician Harold Hotelling, "Stability in Competition", The Economic Journal, March 1929, pages 41-57.

[62] The Hotelling paper argues, in fact, the providers will tend to prefer to provide products very much like those of their competitors, even at the cost of leaving some demand only imperfectly satisfied.

[63] See, for example, Joskow, P., "Regulation and Deregulation after 25 Years: Lessons Learned for Research in Industrial Organization", 2004, pages 26-27, available at:

http://econ-www.mit.edu/faculty/download_pdf.php?id=1005.

[64] Alfred E. Kahn, "Whom the Gods would Destroy, or How not to Deregulate", available at http://www.aei.brookings.edu/admin/authorpdfs/page.php?id=112.

[65] Jean-Jacques Laffont, J. Scott Marcus, Patrick Rey, and Jean Tirole, "Internet interconnection and the off-net-cost pricing principle", RAND Journal of Economics, Vol. 34, No. 2, Summer 2003, available at http://www.rje.org/abstracts/abstracts/2003/rje.sum03.Laffont.pdf.

[66] This is not altogether true. My former firm, BBN, operated a commercial RSVP-based network for many years. It was a commercial failure, but not a technical failure.

[67] Those of us who remember international telephone calls routed over satellites are familiar with this phenomenon.

[68] For an introduction to the use of queuing theory in this context, see Chapter 16 of my textbook, Designing Wide Area Networks and Internetworks: A Practical Guide, Addison Wesley, 1999.

[69] The graph was computed using the Pollaczek-Khinchine formula for an M/G/1 queuing model. This implies a Markovian arrival pattern; however, the so-called operational analysis school of queueing theory has demonstrated that the formula can also be derived under greatly relaxed assumptions. A mean packet length of 284 octets is assumed, consistent with observational experience around 2001.

[70] This was, of course, the key root problem in BBN's inability to successfully commercialize its RSVP-based commercial QoS-capable network.

[71] In a classic joke, a child looks for a lost coin under a lamp post, not because he lost it there, but rather because that is where the light is best.

[72] Jeffrey H. Rohlfs, Bandwagon Effects In High-Technology Industries 3 (2001). Much of the discussion in this section derives from Rohlfs's excellent book.

[73] I make this case at much greater length in "Evolving Core Capabilities of the Internet", Journal on Telecommunications and High Technology Law, 2004.

[74] If each of n interconnected networks need to reach agreement with every other network, this implies a need for n(n-1)/2 interconnection agreements. The number of agreements goes up as the square of the number of networks.

[75] The thought here is to provide examples of contractual arrangements that seem to work, but emphatically not to intrude on the ability of commercial service providers to conclude whatever arrangements they might choose.

[76] Including slow rolling, cost-price squeezes, and strategic litigation.

[77] In the absence of regulation, these behaviors can arise quickly and spontaneously. In the United States in the early 1900's, it was a refusal of AT&T to interconnect with competitors that led to the Kingbury Commitment of 1912, and ultimately to the regulation of telecommunications.

[78] Justus Haucap and J. Scott Marcus, "Why Regulate? Lessons from New Zealand", IEEE Communications Magazine, November 2005, available at: http://www.comsoc.org/ci1/Public/2005/nov/ (click on "Regulatory and Policy").

[79] See M. Katz and C. Shapiro (1985), "Network externalities, competition, and compatibility", American Economic Review 75, 424-440.; and J. Farrell and G. Saloner (1985), "Standardization, compatibility and innovation", Rand Journal of Economics 16, 70-

83. Two threads of economic research, one related to standards, the other to interconnection, proceeded in parallel for many years. Only later did the economists realize that the underlying economics were nearly identical.

[80] Jacques Cremer, Patrick Rey, and Jean Tirole, Connectivity in the Commercial Internet, May 1999. Note that the author was a prominent intervener in both cases.

[81] The emergence of "network neutrality" as a hot issue in the United States may reflect a recognition or belief, perhaps not fully understood or articulated, that broadband Internet providers in the U.S. might be approaching this threshold. Whether this is really so remains unclear.

[82] It must, however, be noted that a framework of this type requires some sophistication as regards economics. Moreover, the effectiveness of implementation depends on institutional arrangements that enable economic tests to be applied impartially and transparently.

[83] Nicholas Garnham, "Contradiction, Confusion and Hubris: A Critical Review of European Information Society Policy", available at http://www.encip.org/document/garnham.pdf.

[84] See Haucap, J., and Marcus, J.S., "Why Regulate? Lessons from New Zealand", IEEE Communications Magazine, November 2005, available at: http://www.comsoc.org/ci1/Public/2005/nov/ (click on "Regulatory and Policy").

[85] Ofcom's approach to risk in the assessment of the cost of capital: Final statement, August 18, 2005

[86] Ofcom defines a real option as "… the term given to a possibility to modify a project at a future point." It relates to "… the option for a firm that faces significant demand uncertainty to 'wait and see' how the demand or technology for a new product will evolve before making an investment."

[87] Content in this context should be construed broadly. It is any information that one might possibly access using the Internet. It could be a website, or a movie, or an audio recording.

[88] "The chief executive of AT&T, Edward Whitacre, told Business Week last year that his company (then called SBC Communications) wanted some way to charge major Internet concerns like Google and Vonage for the bandwidth they use. 'What they would like to do is use my pipes free, but I ain't going to let them do that because we have spent this capital and we have to have a return on it,' he said." New York Times, March 8, 2006

[89] FCC, "In the Matter of Madison River Communications, LLC and affiliated companies", available at:

http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-05-543A1.pdf.

[90] The author has no first hand knowledge of this case.

[91] Marcus, J.S., "Is the U.S. Dancing to a Different Drummer?" Communications & Strategies, no. 60, 4th quarter 2005. Available at: http://www.idate.fr/fic/revue_telech/132/CS60%20MARCUS.pdf.

[92] Referred to in this context as international settlement rates.

[93] I am not a neutral party in the matter. I have an ongoing relationship with the Jamaican regulatory authority.

[94] FCC, Modifying the Commission's Process to Avert Harm to U.S. Competition and U.S. Customers Caused by Anticompetitive Conduct, IB Docket No. 05-254, Released: August 15, 2005.

[95] See also the remarks of Thilo Salmon of SIPgate at the recent NGN and Emerging Markets workshop, Koenigswinter, Germany, December 5 2006.

[96] The IP address reflects the topological location. The telephone number implies a geographic location, but that implication will not necessarily be reliable for "nomadic" services, including VoIP.

[97] In this section in particular, I am my own primary source. When I was in industry as a Chief Technology Officer for a major Internet backbone service provider, I was very active in trying to evolve peering arrangements to accommodate QoS.

[98] Mike Muuss, "The Story of the PING Program", available at http://ftp.arl.mil/~mike/ping.html.

[99] See, for instance, http://www.ripe.net/projects/ttm/about.html.

[100] See Next Generation Networks – Future arrangements for access and interconnection, October 24, 2004; and Next Generation Networks: Further consultation, June 30, 2005.

[101] See M. Katz and C. Shapiro (1985), "Network externalities, competition, and compatibility", American Economic Review 75, 424-440.; and J. Farrell and G. Saloner (1985), "Standardization, compatibility and innovation", Rand Journal of Economics 16, 70-83.

[102] If BigCo were to refuse to peer on reasonable terms with any domestic competitors, it is possible that regulatory intervention might be appropriate. There are parallels to the circumstances that pertained in Australia a few years ago, where the government considered it necessary to impose a peering obligation on their historic incumbent.

[103] For example, circuits from a foreign provider to a commercial end user will be longer and more expensive than circuits from BigCo, in general.

[104] These "secondary" peering arrangements will tend to emerge spontaneously, without regulatory intervention. They are already evident, for example, among VoIP providers in the UK.

Exhibit G

Journal on Telecommunications & High Technology Law
2004

The Digital Broadband Migration: Toward a Regulatory Regime for the Internet Age
A symposium co-sponsored by the Journal on Telecommunications and High Technology Law and the Silicon Flatirons Telecommunications Program

***121** EVOLVING CORE CAPABILITIES OF THE INTERNET

J. Scott Marcus [FNa1]

Copyright © 2004 Journal on Telecommunications & Hight Technology Law; J.

Scott Marcus

Abstract

Historically, the Internet has served as an enormous hotbed of innovation. Nonetheless, deployment of a number of potentially beneficial and important Internet capabilities appears to be slowed or stalled for lack of sufficient commercial incentives. The primary concern is with public goods [FN1] where market forces alone might not be sufficient to drive widespread adoption. Timely and relevant examples are drawn primarily from the areas of network security and cybersecurity. How might government identify and prioritize those capabilities where intervention is warranted (if ever)? What actions on the part of industry and government are necessary and appropriate in order to ensure that societally significant problems, including network security and robustness, are addressed in the Internet?

***122** Table of Contents

Abstract 121

Introduction 123

I. Barriers to Adoption 126

A. Transaction Costs 127

B. Network Externalities 128

C. Misalignment of Incentives 130

D. The Time Frame of Risks and Rewards 131

E. The TCP/IP Reference Model 131

F. The End-to-End Principle 136

II. The Technology of DNS Security 139

A. The Domain Name System 139

B. Security Exposures in the DNS 140
C. DNS Security Mechanisms 141
1. Domain Name System Security Extensions 141
2. Secret Key Transaction Authentication for DNS (TSIG) 143
D. Deployment of DNS Security Mechanisms 144
III. Public Policy Alternatives 146
A. Provide Leadership 147
B. Help Industry to Forge a Consensus 148
C. Stimulate Standards Bodies to Focus on Relevant Problems..... 149
D. Collect Relevant Statistics..... 150
E. Provide "Seed Money" for Research and for Interoperability Testing ... 151
F. Support Desired Functionality in Products and Services Through Government's Own Purchasing Preferences 152
G. Fund the Deployment of Desired Capabilities 155
H. Mandate Use of Desired Services..... 156
I. Adoption of the Metric System - A Sobering Case Study 157
J. Funding for the Early Internet - A Happier Case Study 159
IV. Concluding Remarks 160

***123** Introduction

Many have argued that the Internet is far more hospitable to innovation than the traditional public switched telephone network (PSTN). [FN2] Not so long ago, it seemed that all things were possible in the free-wheeling entrepreneurial and unregulated culture of the Internet. Nonetheless, it now appears that many seemingly promising innovations have languished in recent years. Is it possible that the Internet is hospitable to some innovations, but not to others? Is it possible that pure free market mechanisms will fall short in cases that are of vital importance to society at large? Might there be a role for government to play in promoting societally valuable goals that the market alone would not achieve? If so, what measures are available to government or industry to attempt to promote adoption of important and beneficial innovations?

One federal report, the draft version of The National Strategy to Secure Cyberspace, posed the key question succinctly: "How can government, industry, and academia address issues important and beneficial to owners and operators of cyberspace but for which no one group has adequate incentive to act?" [FN3] The final version of that same report offers an answer: "The government should play a role when private efforts break down due to a need for coordination or a lack of proper incentives." [FN4]

***124** A particular concern here is with public goods. The Economist defines public goods as:

Things that can be consumed by everybody in a society, or nobody at all. They have three characteristics. They are:

• non-rival - one person consuming them does not stop another person consuming them;

• non-excludable - if one person can consume them, it is impossible to stop another person consuming them;

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

• non-rejectable - people cannot choose not to consume them even if they want to.

• Examples include clean air, a national defense system and the judiciary. The combination of non-rivalry and non-excludability means that it can be hard to get people to pay to consume them, so they might not be provided at all if left to market forces . . . . [FN5] Most of the examples in this paper are drawn from the fields of network security and cybersecurity. In the aftermath of the events of September 11, 2001, there is a widespread recognition of the need to enhance the robustness and security of the Internet. Many security exposures exist. Techniques are available to prevent or at least mitigate the impact of the exploitation of certain of the known exposures; however, in certain instances, it is not clear that the organizations that would need to make investments to deploy the technologies are motivated to do so. This is especially likely where deployment costs would exceed the quantifiable economic benefits to the organizations that would have to bear those costs.

The Internet is unquestionably one of the greatest technological successes of modern times. Among the many factors that contributed to its success is the end-to-end model, which enables innovation at the edge of the network without changes to the core; and the absence of central control or regulation, which has enabled the Internet to evolve largely through private initiative, without the restrictions of cumbersome governmental oversight. To a large degree, the Internet represents a triumph of unbridled capitalist initiative.

Today, most networking professionals would agree that the Internet would benefit from a number of evolutionary changes - changes which, however, appear not to be forthcoming. In many cases, the technology ***125** seems to be sufficiently straightforward, but deployment is stymied by a constellation of factors, including:

• the lack of sufficient economic drivers;

• the difficulty of achieving consensus among a plethora of stakeholders with interests that are either imperfectly aligned or else not aligned at all; and;

• the inability of government to foster change in an entity that is global in scope, and largely unregulated in most industrialized nations.

In other words, the very factors that fostered the rapid evolution of the Internet in the past may represent impediments to its further evolution. Historically, those Internet features that could be implemented through private initiative at the edge of the network emerged rapidly; those features that now require coordinated changes, and especially changes to the core of the network, are either slow to emerge or are not emerging at all. [FN6] One might now wonder whether the Internet has reached an evolutionary cul-de-sac.

This paper draws on examples associated with network security and cyber security; however, the issue of promoting public goods where market forces would otherwise be insufficient is a much larger topic. The author humbly asks the reader's indulgence as he frenetically jumps back and forth from the general to the more specific.

Readers who are well versed in the technology of the Internet may have an easier time following the issues, but this paper is not primarily about technology; rather, it focuses on the business, economic and regulatory factors that serve either to facilitate or to impede evolution. In any case, with the possible exception of Section II (which the reader could skip without loss of continuity), no prior knowledge beyond that of an intelligent layman is assumed as regards any of these disciplines.

This introduction provided a cursory overview of the issues. Section I provides background on factors that may militate against the deployment of certain kinds of enhancements to Internet functionality: the end-to-end principle, transaction costs, and the economics of network externalities (following the seminal work of Jeffrey Rohlfs). [FN7] Section II provides a brief technical overview of two emerging security ***126** enhancements to the Domain Name Service (DNS), which collectively serve as an example of seemingly desirable security capabilities and the associated deployment challenges. Section III gingerly explores a topic that many in the Internet community will find uncomfortable: whether it is appropriate for government to play a more active role in fostering the further technical evolution of the Internet. Government intervention could be positive; it could be ineffective; or it could be counterproductive. What role, if any, should the U.S. Government play in the future technical evolution of the Internet? Section IV provides brief concluding observations.

I. Barriers to Adoption

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

As part of the process of preparing the National Strategy to Secure Cyberspace, the President's Critical Infrastructure Protection Board (CIPB) convened a group of Internet experts. At a meeting of this group in May 2002, I commended them for their excellent and thoughtful recommendations. [FN8] I noted the importance of their work, and encouraged them to let their colleagues in government know if, as their work proceeded, they encountered difficulties in getting their firms to deploy the recommended facilities.

A moment of embarrassed silence followed. One of the attendees then timorously put up his hand and said:
Scott, you don't have to wait a year or two to find out whether we are having problems getting this stuff deployed. We already know the answer. There is nothing new in these reports. All of this has been known for years. If we were able to craft business cases for our management, all of this would have been done long ago. No one who has dealt with these issues in industry should be surprised by this answer. Certain Internet innovations have achieved widespread use with no market intervention, perhaps the most noteworthy being the World Wide Web. A great many other Internet innovations have languished, even though the underlying technology appeared to be sound.

***127** In addition to the DNS security facilities described in this report, similar deployment concerns might be raised about:

• Internet Protocol (IP) version 6 [FN9]

• Differentiated services (DiffServ) [FN10]

• IP multicast

• Operational tools and protocol enhancements to enhance the security of BGP-4 routing protocols Engineers tend to conceptualize these deployment delays in terms of engineering concerns, such as incomplete protocol specifications, immature protocol software implementations, and insufficient interoperability testing. It may well be that these engineering problems are symptomatic of deeper business and economic impediments that militate against deployment and use of certain kinds of innovations in the Internet today.

This section of the paper discusses a constellation of economic factors that impede deployment of certain kinds of Internet facilities. The detailed interplay among these factors, and perhaps among other factors not considered here, may vary from one service to the next, but much of the observed behavior can apparently be explained by a small number of underlying economic factors.

A. Transaction Costs

Transaction costs are the economic costs associated with effecting a transaction. [FN11] Some transactions involve far higher transaction costs than others. If a customer buys a candy bar in a luncheonette, she typically hands the cashier some money, receives her change, and walks out the door with the desired item. Transaction costs are low. If that customer ***128** purchases by credit card, the merchant pays a fee for the use of that credit card - transaction costs are higher. If a person buys or sells a house, transaction costs (broker's fees, loan initiation, and various fees) might consume a hefty 5-10% of the value of the transaction.

Transaction costs thus represent sand in the gears, a form of economic friction. Where a large number of parties must independently come to terms with one another on a single transaction, and particularly where those terms require substantial discussion or negotiation, transaction costs will tend to be very high.

High transaction costs cut into the surplus (the degree to which the value to a purchaser exceeds the cost) associated with a transaction. High transaction costs can literally be prohibitive - they can make the transaction as a whole uneconomic. Those who claim that the Internet is a hotbed of innovation are implicitly arguing that transaction costs to deploy new innovations on the Internet are low. In the pages that follow, this paper suggests that this is true only for certain kinds of innovations.

B. Network Externalities

The value of a network is largely a function of who can be reached over that network. Robert Metcalfe, the co-inventor of the Ethernet Local Area Network, attempted to roughly quantify this in Metcalfe's Law, which claims that the value of a network is roughly proportionate to the square of the number of users. [FN12]

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

Most industries experience economies of scale - bigger is better. Networks, however, are subject to additional effects of scale that go far beyond traditional economies of scale. Every time that someone in North Dakota obtains telephone service for the first time, it enhances the value of everyone's telephone service - there is one more person who can be reached by phone. Economists refer to these effects as network externalities, or informally as bandwagon effects.

For a product or service subject to substantial network externalities, nothing succeeds like success. One of the most common examples of a bandwagon effect is the competitive clash of two videocassette standards, VHS and Betamax. At a technical level, neither had a decisive advantage over the other, and for a time they coexisted in the marketplace. Over time, VHS acquired more customers. As a result, studios developed more programming in the VHS format. Consumers with Betamax ***129** equipment found less and less of interest in rental stores, and eventually nothing at all. "Eventually, all consumers - even those who preferred Beta[max]'s picture quality . . . - had no choice but to get on the VHS bandwagon." [FN13]

In some instances, network externalities manifest themselves by way of direct interactions with other users of the same network. In others, the bandwagon effects relate to complementary upstream or downstream industries, as was the case with VHS and Betamax (the player was valuable only if extensive content was available to play on it). These complementarities often lead to the classic "chicken and egg" problem, where two vertically related industries cannot succeed unless both are launched at once.

In a bandwagon marketplace, multiple stable equilibria are usually possible, and these equilibria can differ greatly. Rohlfs defines the initial user set as comprising "all individual entities . . . that can justify purchasing the service, even if no others purchase it." [FN14] If the demand for the service is enhanced by being taken up by the initial user set, then additional users will acquire the service until a higher equilibrium is reached, the demand-based equilibrium user set. The level of usage that is societally optimal, the maximum equilibrium set, may be much larger than the demand-based equilibrium user set. [FN15]

Unfortunately, "ordinary demand adjustments do not provide a path to the optimum." [FN16] Achieving the maximum equilibrium set often requires "supply-side activities or government intervention." [FN17]

New technology products and services have to get over an initial "hump" in order to reach critical mass. Different high-technology industries have achieved critical mass in different ways. Large numbers of videocassette recorders (VCRs) were sold to time-shift television programs on a stand-alone basis; subsequently, these VCRs established the necessary preconditions for the videocassette rental business that today represents the primary use of the VCR. [FN18] For CD players, necessary complementary products became available due to vertical integration - the same firms that were manufacturing CD players (Phillips and Sony) had significant ownership interests in producers of recorded music. [FN19] For black and white television, industry convergence on the National Television Standards Committee (NTSC) technical ***130** standard, coupled with its rapid adoption by the FCC, played a large role in overcoming the initial start-up problem. [FN20]

C. Misalignment of Incentives

In a largely unregulated, market-based system, firms make business decisions based on anticipated costs and benefits. Any decision to change a firm's existing operating environment will entail initial costs. If the firm is to incur those costs, it must believe that there will be corresponding benefits that exceed those costs.

A recent report by the Institute for Infrastructure Protection (I3P) describes the dilemma:

In a market-based economic system, it is not surprising that the market for IT and cyber security products defines the state of cyber security. Two closely related questions appear to drive decisions on how security products and services are acquired and used: (1) what are the cyber security risks to the enterprise and how do they fit into the overall risk equation of a company, and (2) what is the value of cyber security - how much financial benefit it provides. There are no clear answers to these questions. [FN21] Features that constitute public goods (such as enhancements to network security) do not in general reduce recurring operating costs, so the benefits must come from somewhere else. Many organizations find it difficult to justify these expenditures for one or more of a number of reasons. Notably, the benefits may be difficult or impossible to quantify, [FN22] or whatever benefits exist may accrue to a party or parties other than the firm that must make the investments. Collectively, these two factors mean that the organization is unlikely to be motivated to make the investment.

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

***131** D. The Time Frame of Risks and Rewards

Après moi, le déluge! (After me, the flood!) [FN23] Firms fund business cases where the expected return exceeds the expected investment within some defined period of time.

Many cyber vulnerabilities relate to potential exploits that have very high cost, but very low probability of occurrence. These are "thirty year flood" events. Firms may resist funding solutions to thirty year flood problems for some combination of reasons, including:

- The business case takes too many years to prove in;
- The risks are too speculative, and thus too difficult to quantify;
- The risks are born primarily by their insurers, or possibly by the government;
- They may believe, rightly or wrongly, that even if the event takes place, they are unlikely to be viewed as negligent if their competitors were similarly unprepared;
- The current managers may consider it unlikely that the event will happen while they are still with the firm. They bequeath the problem, if indeed it proves to be a problem, to their successors.

E. The TCP/IP Reference Model

The underlying architecture of the Internet has significant implications for the transaction costs associated with the deployment of new capabilities. This part of the paper describes the architecture of the Internet in order to motivate the discussion of the economics associated with the end-to-end principle that appears in the subsequent section.

Perhaps the most significant advance of the past thirty years or so in data networking is the advent of layered network architectures. A layered network architecture breaks the functions of a data network up into functional layers, each of which communicates with its peer layers in other communicating systems, while deriving services from the layer ***132** beneath. This layering helps insulate one layer from another, providing many benefits - a topic we return to later in this section of the paper.

The TCP/IP protocol family, or protocol suite, is the preeminent example today of such a layered network architecture. [FN24] The TCP/IP protocol suite is based on a conceptual model that characterizes the communications hardware and software implemented within a single communicating system - for instance, the personal computer (PC) on your desk - as being comprised of a protocol stack containing multiple layers (see Figure 1). [FN25]

Levels 1 and 2, the Physical and Data Link Layers respectively, represent the realization of the "wire" over which communication takes place and the management of that wire. For instance, the Data Link Layer might determine which of several computers is authorized to transmit data over a particular local area network (LAN) at a particular instant in time.

Level 3, the Network Layer, forwards data from one interconnected network to the next. For the Internet, the Network Layer is the Internet Protocol (IP), which independently routes and forwards small units of data (datagrams).

Level 4, the Transport Layer, processes those datagrams and provides them to whichever application needs them, in the form that the application requires. For the Internet, the Transmission Control Protocol (TCP) supports applications that need a clean and reliable stream of data with no omissions or duplicates. The User Datagram Protocol (UDP) represents an alternative Transport Layer protocol that supports applications that do not require the tidy delivery that TCP provides. E-mail uses TCP, while Voice over IP (VoIP) uses UDP.

***133** Figure 1 Protocol layers in the OSI / Internet Reference Model

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

Level 5, the Application Layer, performs useful work visible to the end user, such as the browser or e-mail client (SMTP, HTTP) on your PC.

In this reference model, a layer logically interacts with its peer in a communicating system (see Figure 2). Thus, an Application Layer, such as the web browser in your PC, communicates with its peer process, a web server in a distant computer.

***134** Figure 2 Peer layers logically interact with one another

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

Each layer within a communicating system implements this logical interaction by requesting services from the next lower layer. Thus, the Application Layer requests data from the Transport Layer. In doing so, it uses an interface that intentionally hides the details of how the lower layer implements its service. This information hiding is a key beneficial property of a layered network architecture - it enables the implementation of a layer to change without impacting the layers above or below.

***135** Figure 3 Logical and physical interactions between network protocol layers

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

Figure 3 shows the relationship between logical and physical interactions in the Internet layered network architecture. It also adds another element to our understanding - a router, which is a device that exists solely to forward traffic in the Internet.

The information hiding property of a layered network architecture facilitates technical innovation over time. It also enables network applications to be written once to operate over any underlying transmission technology, or combination of technologies, thus simplifying the application creator's job. Conversely, the creator of a new transmission technology need only ensure that adequate interfaces exist to enable upper layers of the network to use the new communications layer - there is no need to make network applications specifically aware of a new underlying transmission technology. Phrased differently, a new network application will work with existing networks, and no changes are ***136** needed to underlying network transmission technologies. A new network transmission technology will work with existing networks, and no changes will be needed to existing applications. These properties greatly simplify the evolution of the network over time, and thereby reduce the transaction costs associated with network evolution.

F. The End-to-End Principle

In the early Eighties, a number of distinguished computer scientists at MIT propounded the end-to-end principle. [FN26] They noted that certain communications capabilities were most appropriately associated, not with the underlying network, but rather with the application that used the network. End-to-end reliability of transmission, for instance, could truly be assured only at the end points themselves. They further argued that, if the function could only be correctly implemented in the end points of the network, that it was a bad idea to also implement these functions in intermediate systems--doing so introduced not only inefficiencies, but also an increased possibility of error. Internet engineers have generally accepted the end-to-end principle as a basic tenet of network design. Moreover, they have sometimes advanced the further argument that the end-to-end principle fosters the evolution of the Internet, in that it enables new applications to be developed at the edges of the network, without disrupting the underlying core. [FN27]

There is much to be said for this view. For example, the creation of the World Wide Web initially depended primarily on the creation of a browser that could read and interpret existing file formats, and secondarily on servers for HTTP. No prerequisite changes were needed to the underlying TCP/IP protocols, the IP addressing system, or the DNS--these already provided the necessary support. This absence of prerequisite changes in turn reduced the number of parties that had to change their infrastructure - no action was required, for instance, on the part of Internet Service Providers (ISPs). By reducing the number of parties who must act in order to implement a particular change to the Internet, the end-to-end principle reduces the transaction costs associated with the development of new applications, thus fostering the continuing evolution of the Internet. [FN28]

***137** More recently, a revisionist scholar, Christian Sandvig, has called this view into question. [FN29] He notes that this interpretation of the end-to-end principle presupposes that the underlying network already provides all of the functionality that will ever be necessary or desirable. In fact, it is difficult to know the impact of "missing" functionality - people develop applications to fit the functionality that is already available. Nobody takes the time to develop the applications that would have failed due to insufficient support in the underlying network; consequently, there is no obvious "graveyard" of failed applications.

Thus, while the end-to-end principle may tend to facilitate the development of new data networking applications (based in the Transport thru Application Layers of the familiar OSI Reference Model, [FN30] as described earlier in this paper), [FN31] it does nothing to foster the evolution of the underlying functionality associated with the Network Layer and below.

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

As it happens, this same OSI Reference Model has largely succeeded in decoupling and simplifying the evolution of its lowest layers. Below the Network Layer - which for TCP/IP is the Internet Protocol - datagrams can be transmitted over any Data Link Layer that is known to two systems that are topologically [FN32] adjacent. This is so because the lowest layers, the Physical and Data Link Layers, operate on a point-to-point basis.

Some years ago, the Dutch logician Edsgar Dijkstra conceived the notion of structured programming. [FN33] By a clean nesting of logical functionality, it was possible to contain the impact of changes to a program to a defined scope of statements within the program. This greatly enhanced the reliability of programs, and made it much easier to evolve programs (because a change in one part of the program was unlikely to cause unexpected and unpredictable adverse impact somewhere else).

A similar evolution took place for database management systems - by segregating functionality into various schemas, and hiding unnecessary details about how those schemas implemented their ***138** respective functions, the database systems fostered greater reliability and ongoing functional evolution.

The OSI Reference Model attempted to apply similar principles to data networks. The functionality of the network was broken down into seven functional layers (five for the TCP/IP world). The upper layers were associated with the application, the lower layers with the transmission mechanism. Each layer communicated with its peer layer in another communicating system; however, each effectuated this communication by requesting services from the layer beneath it. A layer never needed to know how the underlying layer provided the functionality.

There is no need for the entire Internet to understand any particular Data Link protocol mechanism. A given system that participates in the Internet need only understand those Data Link protocols whereby it communicates with the systems with which it maintains direct point-to-point communications. These systems could be said to be topologically adjacent.

These properties provide a decoupling for the lower layers of the OSI Reference Model that is very similar in effect to that which the end-to-end principle provides for the upper layers. New applications can be implemented as communicating processes in any two cooperating systems. Likewise, new transmission facilities at the Data Link Layer and below can be implemented in any two adjacent cooperating systems. In both cases, the transaction costs associated with deployment are bounded.

All of this breaks down for the Network Layer, IP. IP provides global connectivity and interoperability for the Internet. There are, of course, ways to evolve the IP functionality of the Internet, but these tend to be complex. There is no assurance that a change made between a pair of systems will have no impact on other systems. There is no inherent mechanism for information hiding within the IP Layer. Any functional evolution must be orchestrated with exquisite caution, because there is no guarantee that the unintended consequences of a given change will be limited.

In sum, technology evolution tends to be complex and expensive for the IP Layer, and also for certain other elements of the Internet that are global in scope. Since the transaction costs associated with evolutionary change of these elements are high, the benefits of any proposed evolutionary change would have to be correspondingly high - otherwise, the deployment of the proposed change is likely to stall for lack of a sufficiently compelling business case.

***139** II. The Technology of DNS Security

There are a wide variety of Internet facilities that might logically fall within the scope of this discussion. In order to motivate the discussion, we focus on a specific constellation of potential Internet security features associated with the DNS.

This paper does not attempt to argue whether any particular Internet security service is in some sense essential. Rather, the intent is to provide background on the rationale of a particular Internet service whose relatively slow deployment might in some sense be emblematic of a broader issue, to assume arguendo that there were some pressing requirement for deployment of that service, and then to pose the question: What impediments to deployment are visible today, and what further impediments might we anticipate in the future? By conducting this thought exercise, we come to a better understanding of the challenges that any feature of this type is likely to encounter.

In this sense, DNS security serves merely as a plausible proxy for any of the Internet-based services that we might have considered.

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

A. The Domain Name System

The DNS is the primary mechanism whereby names, such as www.fcc.gov, are mapped to Internet addresses, such as 192.104.54.3. The DNS has other mapping or directory functions as well. [FN34]

A DNS client, which might reside in your PC, initiates a DNS request to determine the IP address of www.fcc.gov. The request might be sent to a DNS server maintained by a company or by an ISP, the firm that provides access to the Internet.

The DNS is usually thought of as representing a logical tree structure. The root of that tree is comprised of thirteen groups of DNS servers in the United States, Europe and Asia. [FN35] Below the root are other groups of servers associated with Top Level Domains (TLDs), which are ***140** associated with the rightmost portion of a domain name [FN36] - for example, .com, .org, or . gov. The servers responsible for .gov provide in turn pointers to the next level down, including servers responsible for .fcc.gov.

This tree structure facilitates delegation of authority.

B. Security Exposures in the DNS

The opening word was inscribed on the archway all the time! The translation should have been: Say 'Friend' and enter. I had only to speak the Elvish word for friend and the doors opened. Quite simple. Too simple for a learned loremaster in these suspicious days. Those were happier times. [FN37] The DNS was designed in happier times, with little or no regard for security concerns. [FN38] When a DNS request is transmitted, there is no assurance that the response came from the desired DNS server, nor that the information provided was valid.

If a malefactor (who somehow had the ability to eavesdrop on DNS requests for the address of www.fcc.gov) wished to subvert the FCC's web site, they would not need to hack www.fcc.gov; they could instead create their own bogus site, and respond to DNS requests with the IP address of the bogus site. They might not even have to block legitimate DNS responses; it would be sufficient to respond faster than the legitimate DNS servers. Users accessing the bogus site would presume it to be the real one. There are countless variants on this scenario. Most of them depend on one of several underlying exposures: [FN39]

***141** • There is no authentication of the DNS server, i.e. no assurance that the server is who it purports to be;

• There is no assured integrity of the DNS response, i.e. no assurance that the message received is the same as that which was sent;

• There is no assurance that the data maintained by the DNS server was not somehow maliciously modified on the server before being sent. There is in any event no assurance that the data is correct;

• Because the DNS is a logical tree, any compromise potentially impacts everything below that point in the DNS tree. There is also concern that malefactors might attempt to cripple large portions of the Internet by launching Distributed Denial of Service (DDoS) attacks against key DNS servers, preventing users from reaching DNS servers. If users cannot resolve certain domain names, then to all intents and purposes they are unable to use the Internet to access those computers. An attack that was launched on October 21, 2002 received considerable media attention. All indications are that the October 21 attacks had minimal impact; nonetheless, the attacks demonstrated that denial of service is a real threat whose impact should not be underestimated.

C. DNS Security Mechanisms

The Internet community has been aware of these security exposures for many years. A number of responses have been developed within the Internet Engineering Task Force (IETF), the relevant standards body. Some of these are potentially more effective than others.

An exhaustive description of these systems is beyond the scope of this paper. The reader who desires more detail should consult the relevant Internet Request for Comments (RFC) documents. I provide a very brief summary here.

1. Domain Name System Security Extensions

The primary response to these security exposures has been the development of a series of specifications for Domain Name Security Extensions, [FN40] notably RFC 2535, that are sometimes termed DNS Security Extensions (DNSSEC). [FN41]

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

***142** RFC 2535 provides for the storage of public cryptographic keys as a new DNS resource record. Keys are used both to authenticate the data's origin, and to assure the integrity of an RRset (a set of DNS resource records).

The authentication mechanism depends on the establishment of a chain of trust. The chain flows from the root of the DNS system (or from some other point in the DNS tree that is by convention assumed to be trustworthy) down to individual DNS leaf entries. The intent is that DNS servers would intrinsically and reliably be aware of the key for the root zone, and would follow trusted and authenticated entries through each level of the DNS tree in order to reach the correct leaf. [FN42]

The creators of RFC 2535 were also concerned about the possible exploitation of negative information in the DNS - responses erroneously claiming that a domain name does not exist. Given that the domain name space is sparse, merely signing the entries that are present would not necessarily prove that a domain name did not exist. RFC 2535 as amended addresses this by providing for an NSEC resource record [FN43] which points to the next valid domain name in what we can loosely term alphabetical order.

RFC 2535 is currently an IETF Proposed Standard. This means that it "is generally stable, has resolved known design choices, is believed to be well-understood, has received significant community review, and appears to enjoy enough community interest to be considered valuable." [FN44] ***143** At the same time, early operational tests have raised questions about a number of important protocol details. [FN45]

RFC 2535 provides for a very comprehensive any-to-any security mechanism, but it is operationally and computationally relatively expensive. There is a natural tendency to focus solely on the incremental cost of hardware and software, but the relevant deployment costs also include training; deployment planning, testing and staging; and ongoing operational complexity and associated incremental expense. Initial generation of public/private key pairs is computationally intensive, as is periodic or episodic re-signing of a DNS zone. Validation of signatures by means of public key cryptography is also computationally intensive - far more so than private key cryptography. The use of RFC 2535 increases the length of DNS responses, and greatly increases the size of the DNS database. [FN46] Ultimately, the cost of increased computational power and server storage may be less important than the incremental expense associated with a substantial increase in operational complexity - ensuring the secrecy of the private keys, and effecting re-signing without breaking the chain of trust are just a few examples. [FN47]

2. Secret Key Transaction Authentication for DNS (TSIG)

A second response has been the use of TSIG to validate, for example, zone transfers [FN48] (the transfer en masse of a possibly large ***144** volume DNS data). [FN49] TSIG serves to verify the origin and authenticity of the DNS data.

TSIG dynamically computes a cryptographic hash in response to a specific DNS request, using the well-known HMAC-MD5 algorithm.

TSIG is felt to be a reasonably mature technology. TSIG depends on a cryptographic signature based on secret keys, and thus depends on the sender and the receiver possessing a shared secret. As TSIG does not provide a key distribution mechanism, it would become unwieldy [FN50] if used to mutually authenticate a large number of systems; however, only a small number of systems typically need to perform (for instance) DNS zone transfers to one another for any particular zone, so TSIG works well enough for its intended purpose.

In comparison with RFC 2535 DNSSEC, TSIG entails far less computational overhead, and does not increase the size of the DNS database. Lewis describes TSIG as less scalable but more efficient than RFC 2535 DNSSEC. [FN51] TSIG provides for authentication and integrity of the data transmitted from the point where it leaves the transmitting server, but it does not authenticate the source data (which may have been compromised in the sending server prior to being transmitted) - in other words, TSIG does not provide full object security. [FN52]

D. Deployment of DNS Security Mechanisms

A number of trial deployments of RFC 2535 DNSSEC have taken place [FN53], but on the whole the system is not in production deployment.

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

In a review undertaken by the IETF in December, 2000, Edward Lewis notes that "[i]n 1999 and 2000, more than a half dozen workshops have been held to test the concepts and the earliest versions of implementations. But to date, DNSSEC is not in common use. The current collective wisdom is that DNSSEC is 1) important, 2) a ***145** buzzword, 3) hard, 4) immature." [FN54] For RFC 2535 DNSSEC, this is hardly surprising. As previously noted, the true costs of deployment are high. [FN55]

In addition, RFC 2535 DNSSEC appears to suffer from many of the characteristics that, as noted in Section I of this paper, potentially complicate deployment. It is not clear that consumers are willing to pay any premium for DNS security; [FN56] given that implementation costs (largely in the form of operational complexity) are significant, those who must invest to deploy the technology will find it difficult or impossible to craft a clear business case. RFC 2535 DNSSEC is strongly influenced by network externality effects - RFC 2535 DNSSEC would be far more valuable to consumers when it is widely deployed than it is today, or even than it would be if it were in modest production deployment. Moreover, because the system depends on a chain of trust, RFC 2535 DNSSEC is of limited value until those chains are established all the way from the DNS root to the PC on the consumer's desk without breaks. [FN57] As all of this implicitly requires the cooperation of many independent parties, the economic transaction costs of a comprehensive deployment would tend to be high. [FN58]

By contrast, indications are that TSIG is deployable today for zone transfers. Per RFC 3130, ". . . one component of DNSSEC, TSIG, is more advanced that the others. Use of TSIG to protect zone transfers is already matured to the 'really good idea to do stage' even if other elements of DNSSEC are not." [FN59]

Based on the discussion of transaction costs earlier in this paper, this is not surprising. The decision to deploy TSIG concerns only a pair (or a small number) of communicating systems, and in most cases a business relationship already exists between the operators of these systems. Thus, transaction costs to deploy are low, and, as we have seen, ongoing costs for computation and storage are also modest. [FN60]

***146** III. Public Policy Alternatives

To the extent that necessary infrastructure enhancements may not be deployed in the absence of intervention, what is the appropriate role for government?

As we have seen, there is no assurance that industry would deploy a service such as secure DNS based solely on commercial incentives, even assuming the best of intentions on the part of all participants. To the extent that services of this type might be important to the security and robustness of the Internet in the United States, this should be cause for concern.

What role should government play in fostering deployment of Internet capabilities where market forces alone might not suffice? How might government identify and prioritize those capabilities where intervention is warranted (if ever)? For such Internet capabilities as we might deem to be vital, what steps are available to private parties and to the U.S. Government to encourage deployment? Which are likely to be most effective? Which are likely to be least intrusive, and least likely to introduce market distortions?

Most of what we have to say in this section of the paper is not limited to DNS security, and for that matter is not limited solely to cyber security issues. The challenge of promoting the deployment of public goods that provide benefits to the public, but where deployment may not be warranted based solely by the workings of the marketplace, comes up in a great many contexts.

Among the options worth considering by government as a means of fostering deployment of societally valuable services where market incentives might not otherwise suffice are:

1. Provide leadership.

2. Help industry to forge a consensus.

3. Stimulate standards bodies to focus on relevant problems.

4. Collect relevant statistics.

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

5. Provide "seed money" for research and for interoperability testing.

6. Support desired functionality in products and services through government's own purchasing preferences.

7. Fund the deployment of desired capabilities.

8. Mandate use of desired services. ***147** An important and overarching consideration is that market intervention should be avoided wherever possible, and kept to a minimum where absolutely necessary. The Communications Act states unambiguously that "[i]t is the policy of the United States . . . to preserve the vibrant and competitive free market that presently exists for the Internet and other interactive computer services, unfettered by Federal or State regulation." [FN61] Henry David Thoreau stated it more tersely: "That government is best which governs least." [FN62]

For a somewhat more expansive comment, we turn to a recent study from the Computer Science and Technology Board ("CSTB") of the National Research Council of the National Academies:

[A]ppropriate market mechanisms could be more successful than direct regulation in improving the security of the nation's IT infrastructure, even though the market has largely failed to provide sufficient incentives for the private sector to take adequate action with respect to information and network security. The challenge for public policy is to ensure that those appropriate market mechanisms develop. How to deal constructively with prevailing market dynamics has been an enduring challenge for government, which has attempted a variety of programs aimed at stimulating supply and demand but which has yet to arrive at an approach with significant impact. Nevertheless, the committee believes that public policy can have an important influence on the environment in which nongovernment organizations live up to their responsibilities for security. [FN63] We now discuss the alternative government options in turn, starting with those that are least intrusive.

A. Provide Leadership

There may be a tendency to overlook the simplest and least intrusive form by which government can seek to foster change: Simply articulating that change is necessary.

It is perhaps counterintuitive that exercise of "the bully pulpit" alone should be sufficient to influence the behavior of industry participants and ***148** other private citizens, [FN64] but there is no question that the simple exercise of government leadership has sometimes driven important change.

Leadership in this sense - sometimes referred to as "jawboning" - is more likely to be most effective where some of the following factors hold:

• Government has succeeded in articulating a clear goal that has broad public support.

• The costs associated with doing as the government requests are small (e.g., within the range of discretionary spending of a senior or chief executive).

• The organization that must act needs to curry the favor of the relevant government agency.

B. Help Industry to Forge a Consensus

The U.S. Government frequently provides fora for discussion in order to help industry to reach consensus. The President's Critical Infrastructure Protection Board (CIPB) did so in meeting with the Internet community in the course of preparing the National Strategy to Secure Cyberspace. [FN65]

Analogously, the FCC encourages the communications industry to work together to enhance overall network robustness through the Network Reliability and Interoperability Council (NRIC). NRIC operates under the Federal Advisory Council Act (FACA). As a FACA, the NRIC provides advice to the FCC; further, NRIC often provides guidance regarding best practices to U.S. industry.

In some instances, this consensus could be expressed as a document or guideline prepared by the participants and embodying industry best practices. FACAs often take this approach.

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

Adhering to industry best practices, as defined by a body such as the NRIC, may also serve to reduce a firm's legal liability to possible allegations of negligence. [FN66] This form of government participation is ***149** generally viewed as positive by industry and by the broader community. It provides government with the opportunity to offer leadership in a minimally intrusive way.

This form of government participation provides industry with an additional benefit. Companies that routinely compete in the marketplace are understandably uncomfortable meeting to discuss joint action, for fear that their discussions could be misconstrued as being anticompetitive. To the extent that the U.S. Government calls firms together to discuss specific issues in the public interest, antitrust concerns tend to be mitigated. [FN67]

C. Stimulate Standards Bodies to Focus on Relevant Problems

One form of industry consensus is embodied in the standards process. As described above, government could play a role in helping industry to agree on a standard. If appropriate, government could perhaps reinforce this result by encouraging the relevant standards body or bodies to officially adopt a standard reflecting that consensus.

In general, government would look to industry to develop solutions for the standards process. Government is not well equipped to pick winners and losers.

For some standards bodies, notably including the International Telecommunications Union (ITU), formal U.S. Government advocacy can play a crucial role in achieving adoption of a standard.

The Internet Engineering Task Force (IETF) is the primary standards body for the Internet. By long-standing tradition, the IETF expects standards participants to present their views as an individual expert, rather than those of the organizations that they represent. The U.S. Government thus plays no formal role in the IETF. Even in this case, however, government can when appropriate facilitate the standards process by supporting research and interoperability testing and by identifying problem areas where it appears that the public interest would be well served by a standards-based solution.

***150** D. Collect Relevant Statistics

In a competitive communications industry, industry participants will have data about their own experiences, but no single industry participant will necessarily have a global view. [FN68]

Government can collect data where appropriate to identify problems, to determine their magnitude, and to provide a basis on which to evaluate potential solutions.

In determining whether to do so, it would invariably be necessary to balance several conflicting objectives. There may be compelling public interest reasons for gathering certain kinds of information; however, collecting that information represents a regulatory burden on the companies involved. That burden should be avoided where possible, and minimized where the data are truly needed.

Another tension of objectives relates to the sensitivity of data gathered. The public has a right to know information held by the Government, as embodied in the Freedom of Information Act (FOIA) and also by various state "sunshine" acts. At the same time industry participants have a legitimate interest in protecting competitively sensitive information, and in preserving the privacy of their customers. Often, these conflicting demands have been reconciled by having a third party anonymize data before providing it to the Government. [FN69]

There are specific exemptions from FOIA that address specific needs. One recent report rightly observes that these exemptions provide agencies with substantial ability to shield information of this type from inappropriate disclosure under FOIA; [FN70] however, that knowledge offers little comfort to industry participants, who must consider not only whether government can avoid inappropriate disclosure of their sensitive data, but also whether it will. [FN71]

***151** In those instances where data collection appears warranted in support of some public policy objective, government can work with industry to define the data required, to evaluate necessary safeguards on the dissemination of that information, and then to establish voluntary reporting programs.

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

Mandatory reporting can be appropriate in some circumstances, but only where the need for the data is compelling, where the data to be collected is well and narrowly defined, and where voluntary reporting for some reason is either inappropriate or unsuccessful.

E. Provide "Seed Money" for Research and for Interoperability Testing

For facilities that may benefit the public interest, but not necessarily individual users or industry participants, it may be that no private funding source is motivated to provide initial "seed" money. Certain security services, for instance, may benefit the public at large rather than any particular individual or company.

Public funding (or funding by public interest sources) may be the only practical way to foster development of such capabilities.

Analogous issues exist with interoperability testing. Many network services are useful only to the extent that they are interoperable with their counterparts in other networks. These counterpart services may be implemented independently and in competing products. Absent testing, there is no assurance that these implementations will interoperate correctly.

The government role in such activities is well established and widely accepted. For an example where this approach worked brilliantly, see the discussion of "Funding for the early Internet - a happier case study" later in this paper. Research [FN72] and interoperability testing may, in addition, serve to facilitate the standards process. The IETF will not progress a standard to Draft Standard status until interoperability among independent implementations has been rigorously demonstrated. [FN73]

***152** F. Support Desired Functionality in Products and Services Through Government's Own Purchasing Preferences

To the extent that the U.S. Government is itself a significant user of data networking services, its buying preferences for its own use can serve to influence the evolution of technology.

This represents an interesting proactive lever for change. Industry and the public tend to view this mechanism as legitimate and non-intrusive. It alters the economic incentives of suppliers, but it works with the economic system rather than against it.

This form of intervention may be particularly useful as a means of motivating suppliers (e.g., of software) to include desired functionality with the standard distribution versions of their products.

At the same time, it should not be viewed as a panacea. Government purchasing power may not be sufficient to drive widespread adoption (which is still subject to the economic effects of network externalities of the larger market). [FN74] Consequently, there is always the risk that government will pay a substantial premium in a vain attempt to foster the development and deployment of features and services that, at the end of the day, prove to be of limited utility.

A case in point is the U.S. Government OSI Profile (GOSIP). A massive international standardization effort was in play in the Eighties and into the Nineties on the part of the International Organization for Standardization (ISO) and the Telecommunication Standardization arm of the International Telecommunications Union (ITU-T). [FN75] They were seeking to develop an entire family of data communications protocols, based on principles of Open Systems Interconnection (OSI). The OSI protocols reflected modern concepts of protocol layering, and a full set of applications, including virtual terminal, file transfer, electronic mail, directory, and network management.

It might seem odd in retrospect that the global standards bodies and governments set out to recreate out of whole cloth functionality that already existed. OSI was nominally open to multiple vendors and implementations, but no more so than TCP/IP. Indeed, at the end of ***153** the day, OSI provided no new functionality that users found significant that was not already available under the TCP/IP protocol suite.

Many foreign governments considered TCP/IP to be the creation of the U.S. Department of Defense. Because TCP/IP had not been created by the recognized international standards process, they considered it inappropriate as the basis for a new, global family of communications standards.

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

The U.S. Government attempted to join a global bandwagon forming in favor of OSI. The National Institutes for Standards and Technology (NIST) published GOSIP Version 1 [FN76] in August 1988, and followed a year later with GOSIP Version 2. [FN77] A profile was needed because many of the OSI protocols were so specified as to permit a variety of mutually incompatible possible realizations. [FN78] As of August 1990, Federal agencies were required to acquire OSI products when they required the functionality supplied by the OSI features specified in GOSIP. There was, however, no requirement that Federal agencies procure only GOSIP-compliant implementations for these purposes, nor was there an obligation for Federal agencies to use the GOSIP-compliant implementations that they had thus procured.

OSI protocols had developed what might have seemed to be an unbreakable momentum in the late Eighties. The ISO and CCITT unequivocally backed the protocols, while the Internet standards groups accepted at least an extended period of coexistence between TCP/IP and OSI protocols. [FN79] Digital Equipment Corporation (DEC), at the time a leading computer manufacturer, had committed to implementing OSI communications protocols in DECNET Phase V.

Today, however, OSI protocols serve as little more than a historical curiosity, an interesting footnote. Why is it that OSI protocols failed to achieve broad market acceptance?

Some have argued (and sometimes with surprising vehemence) that government support was the kiss of death for OSI protocols. This seems, however, to miss the point. In particular, it fails to explain the ***154** success of TCP/IP protocols, which by all accounts benefited enormously from substantial support from the U.S. Government.

Others have argued that OSI protocols were cumbersome, and evolved slowly, because they were developed by large committees and because the protocol specification effort took place in advance of implementation. (Internet protocols, by contrast, would never be standardized until independent implementations had been shown to interoperate.) There probably is some truth to this assertion, and it is moreover plausible in terms of what we know of the economics of transaction costs - the need to obtain concurrence of a great many independent parties invariably exacts costs, one way or another. Nonetheless, it is only a part of the answer.

It must also be noted that OSI protocol implementations tended to be significantly more expensive than TCP/IP protocol implementations, not only in terms of purchase price, but also in terms of memory requirements, processing power requirements, and operational complexity. These were certainly factors, but they may not have been decisive.

A simple and sufficient explanation flows from the economic theory of network externalities. TCP/IP implementations were available on most platforms of interest, and the software was inexpensive or free in many cases, unlike OSI implementations. The deployment of OSI protocols at their peak probably never accounted for more than 1-2% of all traffic on the Internet. Users were motivated to use TCP/IP, because most of the content that they wanted to use or view was available in the TCP/IP world, and not in the OSI world. Content providers and application developers were motivated to use TCP/IP, because the majority of their prospective users were TCP/IP users. (Similar factors may have provided Microsoft Windows with an advantage over the Macintosh and, for that matter, VHS with an advantage over Beta, as noted earlier.)

OSI protocols were starting from a position of zero market share. They could not fully supplant TCP/IP protocols unless they replaced all of TCP/IP's functionality; however, TCP/IP began with a huge head start in functionality. Moreover, ongoing investment in new functionality based on the TCP/IP protocols inevitably outstripped that for new OSI functionality by a wide margin. Given that OSI had no compelling inherent advantage over TCP/IP, there was never any means to reverse this trend.

Eventually, the requirement to procure services implementing GOSIP (and its companion standard, the Government Network ***155** Management Profile (GNMP)) [FN80] was lifted. It was presumably recognized that a mandate to procure GOSIP-compliant solutions no longer served a useful purpose. Meanwhile, the U.S. Government had supported the evolution and testing of OSI protocols in many ways, and Federal agencies likely paid more than they otherwise might have to procure functionality that they ultimately did not need and, for the most part, did not use.

G. Fund the Deployment of Desired Capabilities

If deployment of a service is in the public interest, but not in the individual interest of the firms that must deploy it, and if deployment entails significant costs, then those firms have a significant economic disincentive to deploy. In a competitive,

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

deregulated telecommunications marketplace, it is not clear how those firms could recapture their investment.

In those cases, it may be that the only possibility of achieving widespread deployment will be through some combination of subsidizing or funding that deployment as well as any associated incremental operational costs, or possibly by mandating deployment, or both.

The Communications Assistance for Law Enforcement Act (CALEA) is a case in point. [FN81] CALEA establishes carrier obligations in regard to lawful intercept of communications (e.g. wiretap). No telecommunications customer would wish to pay a premium for the privilege of having his or her own communications amenable to wiretap, nor would any carrier have a business incentive to implement the necessary tools and facilities.

As a result, CALEA establishes the Department of Justice Telecommunications Carrier Compliance Fund [FN82] in an effort to "make the carriers whole." This process has not been painless - carriers have argued that the fund does not adequately reimburse them for costs incurred. [FN83]

***156** Government funding for public goods can take any of a number of forms. It can come from general revenues. It can be a distinct fund, as is the case for CALEA. It can also be a separate fund privately managed on behalf of the government, as is the case for universal service.

H. Mandate Use of Desired Services

If functionality were truly deemed to be essential to the public interest, and if market forces were insufficient to ensure its deployment, then it could in principle be appropriate for government to mandate its deployment and use.

For the Internet, there is no obvious historical example; however, there are many examples in the history of the telephone industry in the United States.

One of these is the previously-noted CALEA. CALEA serves both to oblige telecommunications carriers to provide the technical means of achieving lawful intercept (wiretap) and to provide a mechanism for offsetting their costs in doing so. Lawful intercept is a legitimate societal need, but it does not specifically benefit an individual carrier; consequently, it can only be achieved to the extent that government provides the impetus, in this case by means of an explicit mandate.

Other examples of services that might have been unlikely to deploy absent government action include:

• Disabilities access to telecommunications, [FN84]

• Provision of 911 services, and

• Local number portability. [FN85] This is the most intrusive means the government has of driving deployment. For a number of reasons, it should be used sparingly. [FN86]

First, as our experience with GOSIP demonstrates, government's ability to prognosticate is limited. [FN87] If government is to mandate deployment and use, it must be very certain that the functionality in question is truly necessary.

***157** Second, mandating a function will generally have a tendency to distort the relevant market. Wherever possible, market mechanisms should be preferred over mandates, especially unfunded mandates.

Finally, there is the risk that a government mandate might lock the industry into the use of a particular technology long after market forces would otherwise have obsoleted it.

I. Adoption of the Metric System - A Sobering Case Study

In considering the prospects for achieving deployment by means of government actions short of an outright mandate, it is helpful to consider historical precedents. We have already discussed GOSIP. Another example, albeit from a different technological domain, is conversion to the metric system.

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

In 1971, the National Bureau of Standards published a report, A Metric America, [FN88] recommending "[t]hat the Congress, after deciding on a plan for the nation, establish a target date ten years ahead, by which time the U.S. will have become predominantly, though not exclusively, metric. . . ." [FN89]

The benefits of metric conversion were thought to be manifest. Recognizing this, the U.S. Government has undertaken significant efforts over the years to foster adoption of the metric system, [FN90] including the passage of the Metric Conversion Act of 1975 [FN91] and the issuance of Executive Order 12770 [FN92] in 1991. Nonetheless, thirty-two years after the publication of A Metric America, it can hardly be said that the United States has "become predominantly, though not exclusively, metric" .

In A Metric America, the National Bureau of Standards report recognized that the United States had become an isolated island in a metric world, and identified the potential costs associated with that isolation. They also attempted to quantify the costs of conversion, and the potential benefits - largely in terms of global trade and simplified ***158** education. The Metric Conversion Act of 1975 expressed the advantages in unambiguous bread and butter terms:

(3) World trade is increasingly geared towards the metric system of measurement.

(4) Industry in the United States is often at a competitive disadvantage when dealing in international markets because of its nonstandard measurement system, and is sometimes excluded when it is unable to deliver goods which are measured in metric terms.

(5) The inherent simplicity of the metric system of measurement and standardization of weights and measures has led to major cost savings in certain industries which have converted to that system.

(6) The Federal Government has a responsibility to develop procedures and techniques to assist industry, especially small business, as it voluntarily converts to the metric system of measurement.

(7) The metric system of measurement can provide substantial advantages to the Federal Government in its own operations. [FN93] An important collective effect of the Metric Conversion Act and of Executive Order 12770 has been to require that each Federal agency ". . . to the extent economically feasible by the end of the fiscal year 1992, use the metric system of measurement in its procurements, grants, and other business-related activities, except to the extent that such use is impractical or is likely to cause significant inefficiencies or loss of markets to United States firms, such as when foreign competitors are producing competing products in non-metric units."

The Metric Conversion Act also attempts to "seek out ways to increase understanding of the metric system of measurement through educational information and guidance and in Government publications." The Act established a United States Metric Board [FN94] tasked with carrying out "a broad program of planning, coordination, and public education." The Board was to perform extensive public outreach, to "encourage activities of standards organizations," to liaise with foreign governments, to conduct research and surveys, to "collect, analyze, and publish information about the usage of metric measurements," and to "evaluate the costs and benefits of metric usage." Thus, the metric conversion program attempted, to a lesser or greater degree, to employ essentially every tool available to government short of outright deployment funding or an explicit mandate. [FN95]

***159** These efforts undoubtedly had effect, but not as great an effect as was intended. Why was this?

A variety of reasons have been put forward to explain why the metric transition has not made widespread progress in the U.S. in the past. They include lack of national leadership, reluctance to embark on such a change, and the failure of the voluntary effort that began in 1975. The many competing national priorities and the lack of immediate and visible benefit to a transition clearly were factors. There are political, economic, and social reasons to explain the apparent slow progress and reluctance to make the transition. [FN96] It is not the intent of this paper to trivialize or over-simplify what undoubtedly was a very complex process. The key point that the reader should take away from this case study is that, for certain kinds of innovations where economic incentives are not sufficient to motivate their deployment in a free market system, there can be no assurance that government actions short of deployment funding or an explicit mandate will generate substantial deployment.

J. Funding for the Early Internet - A Happier Case Study

In the case of the Internet, by contrast, the historic effects of direct Government funding have in most instances been salutary. The original ARPAnet, the predecessor to the Internet, was funded in the late Sixties by the Advanced Research Projects Agency of the U.S. Department of Defense (DARPA). [FN97]

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

In the early Eighties, DARPA funded the University of California at Berkeley to incorporate TCP/IP protocols into Berkeley UNIX®. [FN98] This effort produced one of the most widely used TCP/IP implementations. Berkeley UNIX was incorporated into an emerging generation of UNIX workstations, thus fostering precisely the network externalities effects that ultimately enabled TCP/IP to prevail in the marketplace.

***160** The U.S. National Science Foundation (NSF) provided initial funding for CSNET as a limited-function network for the academic research community. The NSF then invested an estimated $200 million from 1986 to 1995 to build and operate the NSFNET as a general purpose Internet backbone for the research and education community. [FN99]

Most observers would agree that the modest investments that DARPA and the NSF made in the Internet have collectively been a brilliant success.

IV. Concluding Remarks

On a hasty reading, this paper might be construed as advocating that government take an intemperate, interventionist approach toward the Internet.

What is called for, in the author's view, is a reasoned and balanced approach. Much has been made of the lack of regulation of the Internet. [FN100] Yet the very existence of the Internet is a direct result of a succession of government interventions, many of them highly successful. Among these were the initial funding of the ARPAnet, the FCC's Computer Inquiries (simultaneously deregulating services like the Internet while opening up underlying telecommunications facilities for their use), support for CSNET and the NSFNET, and the funding of TCP/IP protocol implementation in Berkeley UNIX. [FN101] Each of these achieved important and positive results without resorting to a regulatory mandate.

There have also been failures of government intervention. Perhaps the most relevant was the U.S. Government's support of OSI protocols through GOSIP and the GNMP, as described earlier in this paper. That ultimately unsuccessful attempt to use the purchasing power of government to promote global standards that the marketplace had by and large not demanded, likely resulted in significant diversion of attention and waste of resources on the part of both government and industry.

Another example was metric conversion, where the U.S. Government has attempted a combination of practically every conceivable measure short of an outright mandate but has not achieved the widespread deployment that was hoped for.

***161** Government is neither omniscient nor omnipotent. Government could do too little. Government could also do too much. How to know which is which?

Two principles may be useful going forward:

Balance: Government should recognize both the risks of action and those of inaction, and make cautious and deliberate choices.

Minimalism: Government should choose to err in general on the side of less regulation rather than more. Do not attempt a massive intervention where a less intrusive intervention might suffice. Do not intervene at all unless markets have shown themselves to be unable to deliver a socially important outcome.

[FNa1]. Author's current address: Federal Communications Commission (FCC), Office of Strategic Planning and Policy Analysis, 445 12th Street SW, Washington, DC 20554 and can be contacted at smarcus@fcc.gov. The author is affiliated with both the FCC and the European Commission, but the opinions expressed are solely those of the author, and do not necessarily reflect the views of either agency. The author is deeply indebted to his colleagues Richard Hovey and Jeffery Goldthorp, of the FCC; to Scott Bradner, of Harvard University; to Dale Hatfield, Gary Chapman and Andrew Johnson, of the University of Colorado; and to Scott Rose of the National Institute of Standards and Technology for a wealth of helpful and insightful comments.

[FN1]. The Economist, Economics A-Z, Economist.com, available at http:// www.economist.com/research/Economics (last visited May 10, 2004) (adapted from Matthew Bishop, Essential Economics (2004).

[FN2]. Cf. David Isenberg, The Rise of the Stupid Network, Computer Telephony, Aug. 1997, at 16-26, available at http:// www.hyperorg.com/misc/stupidnet.html.

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

[FN3]. The President's Critical Infrastructure Protection Board, The National Strategy to Secure Cyberspace, Draft for Comment 47 (2002), available at http://www.iwar.org.uk/cip/resources/c-strategy-draft [hereinafter Draft National Strategy to Secure Cyberspace].

[FN4]. The President's Critical Infrastructure Protection Board, The National Strategy to Secure Cyberspace 31 (2003), available at http:// www.whitehouse.gov/pcipb/cyberspace_strategy.pdf [hereinafter National Strategy to Secure Cyberspace].

[FN5]. The Economist, supra note 1. They go on to observe that, "public goods are regarded as an example of market failure, and in most countries they are provided at least in part by government and paid for through compulsory taxation." Id.

[FN6]. Cf. Christian Sandvig, Communication Infrastructure and Innovation: The Internet as End-to-End Network that Isn't (Nov. 2002) (unpublished manuscript, available at http://www.cspo.org/nextgen/Sandvig.PDF).

[FN7]. Jeffrey H. Rohlfs, Bandwagon Effects in High-Technology Industries 3 (2001).

[FN8]. For a public summary of their major findings, see Avi Freedman, Akamai Techs., ISP Working Group Internet Vulnerability Summary & Discussion (2002), available at http://www.nanog.org/mtg-0206/avi.html.

[FN9]. The National Telecommunications and Information Administration (NTIA), which is a part of the U.S. Department of Commerce, is currently conducting a Notice of Inquiry regarding IP version 6. Public comments are available at http:// www.ntia.doc.gov/ntiahome/ntiageneral/ipv6/commentsindex.html. The parallels to DNS security are quite striking.

[FN10]. Within the network of a single service provider, differentiated services are readily achievable. In the general, multiple-provider case, there is no significant deployment.

[FN11]. Various definitions exist in the literature. See, e.g., Organization for Economic Cooperation and Development, Transaction Costs and Multifunctionality, available at http:// www1.oecd.org/agr/mf/doc/Transactioncosts32.pdf (last visited May 26, 2004) (citations omitted). It defines transaction costs in this way: "Transaction costs are 'the costs of arranging a contract ex ante and monitoring and enforcing it ex post' ... 'the costs of running the economic system' ... and 'the economic equivalent of friction in physical systems....' " Id. at 2 (citations omitted).

[FN12]. Cf. Andrew Odlyzko, Content is Not King, First Monday, Jan 8, 2001, at http://www.firstmonday.dk/issues/issue6_2/odlyzko/ (arguing that "...Metcalfe's Law does not reflect properly several other important factors that go into determining the value of a network. However, the general thrust of the argument... [is] valid.").

[FN13]. Rohlfs, supra note 7. (The discussion of network externalities that follows draws heavily on Rohlfs's work.).

[FN14]. Id. at 23.

[FN15]. Id. at 24.

[FN16]. Id.

[FN17]. Id.

[FN18]. Id. at Ch. 10.

[FN19]. Rohlfs, supra note 7, at Ch. 9.

[FN20]. Id. at Ch. 12.

[FN21]. Institute for Information Infrastructure Protection, Cyber Security Research and Development Agenda 40 (2003), available at http:// www.thei3p.org/documents/2003_Cyber_Security_RD_Agenda.pdf [hereinafter I3P Report].

[FN22]. Id. at 34-45.

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

Decision makers lack a foundation of data about the current investment and risk levels: metrics that express the costs, benefits, and impacts of security controls from an economic perspective, technical perspective, and risk perspective; and ways to predict the consequences of risk management choices.... Risk assessment and dependency modeling for cyber security remain in an immature state with only little momentum in the marketplace.

Id.

[FN23]. Attributed to Louis XV, king of France from 1715-1774. Some sources instead attribute this quotation to his mistress, Madame de Pompadour.

[FN24]. The evolution of the TCP/IP protocol suite was influenced by earlier layered network architectures, and influenced in turn the subsequent evolution of a number of those network architectures. Among the layered network protocol families that emerged during the Seventies and Eighties were CCITT's X.25, IBM's System Network Architecture (SNA), Digital Equipment Corporation's DECnet, and Xerox Network Systems (XNS). Perhaps the most influential layered network architecture was the Reference Model for Open Systems Interconnection, usually referred to as the OSI Reference Model. The OSI Reference Model was developed jointly by the International Organization for Standardization (ISO) and the ITU/CCITT. The most readable descriptions of the OSI Reference Model appear in Hubert Zimmerman, OSI Reference Model - The ISO Model of Architecture for Open Systems Interconnection, 4 IEEE Transactions on Comm. 425 (1980), and in Andrew Tanenbaum, Computer Networks (Prentice Hall 3d ed. 1996).

[FN25]. Rigid adherence to protocol layering tends to impose a high overhead on protocol software. In reality, TCP/IP implementations often combine layers or take short-cuts as a means of reducing this overhead. See David D. Clark, RFC 0817: Modularity and Efficiency in Protocol Implementation (Internet Engineering Task Force, July 1982), at http://www.ietf.org/rfc.html.

[FN26]. J.H. Saltzer et al., End-to-End Arguments in System Design, in ACM Transactions on Computer Systems 2, 277 (1984), available at http:// web.mit.edu/ Saltzer/www/publications/endtoend/endtoend.pdf.

[FN27]. Isenberg, supra note 2.

[FN28]. For an interesting economic interpretation of the costs and benefits of this flexibility, see Mark Gaynor et al., The Real Options Approach to Standards for Building Network-based Services (2nd IEEE Conference on Standardization and Innovation in Information Technology, Oct. 2001), available at http://people.bu.edu/mgaynor/papers/IEEE-standard-camera.pdf.

[FN29]. Sandvig, supra note 6.

[FN30]. Zimmerman, supra note 24 (the TCP/IP protocol suite that forms the foundation of the Internet broadly follows the OSI Reference Model, but with simplification in the upper layers).

[FN31]. See supra Section I.E.

[FN32]. Topology is the branch of mathematics that deals with the interconnectivity of the vertices and edges that comprise geometric figures, without considering their dimensions. It provides a useful way to visualize communications networks and to express their formal properties.

[FN33]. O.J. Dahl et al., Structured Programming (1972).

[FN34]. The DNS is documented in a series of Requests for Comments (RFC) that were developed by the Internet Engineering Task Force (IETF). The primary references are P.V. Mockapetris, RFC 1034: Domain names - concepts and facilities (Internet Engineering Task Force, Nov. 1, 1987), at http:// www.ietf.org/rfc.html [hereinafter RFC 1034] (updated by RFC 1101, RFC 1183, RFC 1348, RFC 1876, RFC 1982, RFC 2065, RFC 2181, RFC 2308, RFC 2535); and P.V. Mockapetris, RFC 1035: Domain names - Implementation and Specification (Internet Engineering Task Force, Nov. 1, 1987), at http:// www.ietf.org/rfc.html [hereinafter RFC 1035] (updated by RFC 1101, RFC 1183, RFC 1348, RFC 1876, RFC 1982, RFC 1995, RFC 1996, RFC 2065, RFC 2136, RFC 2181, RFC 2137, RFC 2308, RFC 2535, RFC 2845, RFC 3425, RFC 3658). All RFCs are available at http://www.ietf.org/rfc.html.

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

[FN35]. Some of these root servers are now mirrored in multiple locations.

[FN36]. Strictly speaking, we should say the rightmost customarily visible portion of the domain name. The rightmost portion is a period denoting the root itself, which is unnamed; however, this is often omitted by convention.

[FN37]. J.R.R. Tolkien, The Fellowship of the Ring 402 (Ballantine Books 1965).

[FN38]. Cf. I3P Report, supra note 21, at iii ("The information infrastructure, taken as a whole, is not an engineered system.... Security was not a significant consideration at its inception, and security concerns today do not override market pressures for new uses of technology or innovation, in spite of frequent stories of hackers, criminals, and, increasingly, terrorists and nations using or planning to use the information infrastructure as a weapon to harm the United States.").

[FN39]. Cf. D. Atkins & R. Austein, RFC __: Threat Analysis of the Domain Name System (Internet Engineering Task Force, Feb. 2004), at http:// www.ietf.org/internet-drafts/draft-ietf-dnsext-dns-threats-07.txt (work in progress: RFC is in preparation). Atkins and Austein primarily characterize threats as (1) packet interception, (2) ID guessing and query prediction, (3) name games, (4) betrayal by trusted server, and (5) denial of service. Id. Much work has been done over the years to characterize threats to the DNS, notably including Steven Bellovin, Using the Domain Name System for System Break-Ins, USENIX, (Jun. 1995), at http:// www.usenix.org/publications/library/proceedings/security95/bellovin.html.

[FN40]. Donald Eastlake III, RFC 2535: Domain Name System Security Extensions (Internet Engineering Task Force, Mar. 1999), at http:// www.ietf.org/rfc.html (updated by RFC 2931, RFC 3007, RFC 3008, RFC 3090, RFC 3226, RFC 3445, RFC 3597, RFC 3655, RFC 3658) [hereinafter RFC 2535]; Donald Eastlake III, RFC 2541: DNS Security Operational Considerations (Internet Engineering Task Force, Mar. 1999), at http://www.ietf.org/rfc.html [hereinafter RFC 2541].

[FN41]. To avoid confusion, we use the term "RFC 2535 DNSSEC" to refer specifically to RFC 2535 capabilities. Some sources use DNSSEC to refer only to RFC 2535, while others use it to encompass additional capabilities, including TSIG, secure dynamic updates (per RFC 3007), and the CERT resource record (RFC 2538).

[FN42]. This seemingly simple assumption masks a world of complexity. For example, the root signature, like all signatures, should be periodically changed in case it has been somehow compromised, and also to minimize the risk of cryptanalysis. If the key is statically configured in every client, how can it reliably be updated? See RFC 2541, supra note 40. See also RFC 2535, supra note 40, at § 6.2.

[FN43]. In the original RFC 2535, the corresponding RR was referred to an NXT resource record. Based on operational experience, a number of non-backward-compatible changes were made to the DNSSEC protocols, culminating in a renaming of several RRs and renumbering of their code points. See S. Weiler, RFC 3755: Legacy Resolver Compatibility for Delegation Signer (DS) (Internet Engineering Task Force, May 2004), at http://www.ietf.org/rfc.html [hereinafter RFC 3755].

[FN44]. Scott Bradner, RFC 2026: The Internet Standards Process -Revision 3, § 4.1.1 (Internet Engineering Task Force, Oct. 1996), at http:// www.ietf.org/rfc.html [hereinafter RFC 2026].

[FN45]. For more information on this topic, visit Ripe NCC, Deployment of Internet Security Infrastructures, at http://www.ripe.net/disi/ (last visited May 26, 2004).

[FN46]. One source claims that it increases the size of the DNS database by a factor of seven. See Paul Albitz & Cricket Liu, DNS and Bind 308-74 (4th ed. 2001), available at http://www.oreilly.com/catalog/dns4/chapter/ch11.html.

[FN47]. Id. at 374 ("We realize that DNSSEC is a bit, er, daunting. (We nearly fainted the first time we saw it.)").

[FN48]. P. Mockapetris, RFC 1034: Domain Names - Concepts and Facilities § 4.3.5 (Internet Engineering Task Force, Nov. 1987), at http:// www.ietf.org/rfc.html [hereinafter RFC 1034]. RFC 1034, describes DNS zone transfers in this way:

"Part of the job of a zone administrator is to maintain the zones at all of the name servers which are authoritative for the zone. When the inevitable changes are made, they must be distributed to all of the name servers. While this distribution can

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

be accomplished using FTP or some other ad hoc procedure, the preferred method is the zone transfer part of the DNS protocol. The general model of automatic zone transfer or refreshing is that one of the name servers is the master or primary for the zone. Changes are coordinated at the primary, typically by editing a master file for the zone. After editing, the administrator signals the master server to load the new zone. The other non-master or secondary servers for the zone periodically check for changes (at a selectable interval) and obtain new zone copies when changes have been made." Id.

[FN49]. Paul Vixie et al., RFC 2845: Secret Key Transaction Authentication for DNS (TSIG) (Internet Engineering Task Force, May 2000), at http:// www.ietf.org/rfc.html (updated by RFC 3645).

[FN50]. In other words, the two systems participating in a TSIG exchange would have to both know the shared secret through some means other than TSIG itself, since TSIG contains no mechanism for distributing the keys. If the keys are to be transmitted through the Internet, by e-mail for example, they must be protected from disclosure to third parties. All of this adds complexity. Since TSIG is normally used for a bounded set of problems where a trust relationship already exists between two systems, the protocol designers have not felt that this extra complexity was warranted.

[FN51]. See generally Edward Lewis, RFC 3130: Notes from the State-Of-The-Technology: DNSSEC (Internet Engineering Task Force June 2001), at http:// www.ietf.org/rfc.html.

[FN52]. See Paul Vixie et al., supra note 49, at § 6.3; see also Atkins & Austein, supra note 39.

[FN53]. See Lewis, supra note 51; see also RIPE NCC, supra note 45.

[FN54]. Lewis, supra note 51, at § 1.0.

[FN55]. See supra Section II.C.1.

[FN56]. There are also open questions regarding the willingness and ability of consumers to cope with the complexity that DNSSEC implies. Suppose the DNSSEC client software were to notify the consumer that the DNS pointer to a commercial web site such as www.amazon.com had been corrupted. It is not clear what action the consumer should then take, since recovery will generally be beyond the consumer's capabilities. In light of this ambiguity, can the DNSSEC client software provide meaningful and sufficient guidance to the consumer?

[FN57]. DNSSEC will be of no use to the average consumer until and unless it is available in the operating system for the consumer's PC - typically Microsoft Windows™

[FN58]. Some have argued for a more piecemeal, selective approach to deployment, but the DNSSEC standards do not currently embrace this approach.

[FN59]. Lewis, supra note 51.

[FN60]. Unfortunately, the benefits are also modest for the reasons previously noted. The threats that TSIG guards against are generally irrelevant to the consumer mass market.

[FN61]. 47 U.S.C. § 230(b)(2) (2000).

[FN62]. Henry David Thoreau, Civil Disobedience (1849), available at http:// www.cs.indiana.edu/statecraft/civ.dis.html (quotation is sometimes attributed to Thomas Jefferson).

[FN63]. Information Technology for Counterterrorism: Immediate Actions and Future Possibilities 104 (John L. Hennesy et al. eds., 2003) [hereinafter Hennesy et al.].

[FN64]. Cf. I3P Report, supra note 21, at 40 ("Currently, the federal government's approach relies on public-private partnerships and the influence of persuasion; more rigorous analysis needs to be done on the prospects for success of this approach.") (emphasis added).

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

[FN65]. Draft National Strategy to Secure Cyberspace, supra note 3.

[FN66]. Potential tort liability, where a firm might be alleged to have taken less than reasonable care to secure its infrastructure against cyberattacks is an emerging, but still largely undeveloped area of the law. See Critical Information Infrastructure Protection and the Law: An Overview of Key Issues (Cynthia A. Patterson & Stewart D. Personick eds., 2003), available at http://www7.nationalacademies.org/cstb/pub_ciip.html [hereinafter Critical Information Infrastructure Protection and the Law].

[FN67]. As a somewhat related example, the National Strategy to Secure Cyberspace recognizes the importance of establishing mutual assistance agreements to help infrastructure sectors respond to cybersecurity emergencies. See National Strategy to Secure Cyberspace, supra note 4, at 24 (stating that the "[Department of Justice] and the Federal Trade Commission should work with the sectors to address barriers to such cooperation, as appropriate." (emphasis omitted)).

[FN68]. Cf. National Strategy to Secure Cyberspace, supra note 4, at 19 ("There is no synoptic or holistic view of cyberspace. Therefore, there is no panoramic vantage point from which we can see attacks coming or spreading.").

[FN69]. For example, when industry participants provide incident reports to Information Sharing and Analysis Centers (ISACs) operating under PDD-63, the information might be sanitized or anonymized before being shared with other ISAC participants or with the government.

[FN70]. See Critical Information Infrastructure Protection and the Law, supra note 66, at 25-29.

[FN71]. Notably, the Homeland Security Act specifically exempts information about critical infrastructure vulnerabilities provided voluntarily from FOIA obligations. Cf. President's Critical Infrastructure Protection Board, supra note 4, at 25 ("the legislation encourages industry to share information with DHS by ensuring that such voluntarily provided data about threats and vulnerabilities will not be disclosed in a manner that could damage the submitter." This is an area of ongoing concern for the DHS, which is working to "... establish uniform procedures for the receipt, care, and storage... of critical infrastructure information that is voluntarily submitted to the government.").

[FN72]. See President's Critical Infrastructure Protection Board, supra note 4, at 34-35 (explicitly recognizing the importance of prioritizing the Federal research and development agenda and tasking the OSTP with doing so).

[FN73]. Bradner, supra note 44.

[FN74]. Cf. Hennessy et al., supra note 63, at 103 ("the IT sector is one over which the federal government has little leverage. IT sales to the government are a small fraction of the IT sector's overall revenue, and because IT purchasers are generally unwilling to acquire security features at the expense of performance or ease of use, IT vendors have little incentive to include security features at the behest of government alone.").

[FN75]. At the time, this was the International Telephone and Telegraph Consultative Committee (CCITT). See International Telecommunications Union, ITU Overview - History (Feb. 13, 2002), at http:// www.itu.int/aboutitu/overview/history.html.

[FN76]. Approval of Federal Information Processing Standards Publication 146, Government Open Systems Interconnection Profile (GOSIP), 53 Fed. Reg. 32,270, 32,270-02 (Dep't Commerce Aug. 24, 1988).

[FN77]. Proposed Revision of Federal Information Processing Standard (FIPS) 146, G3OSIP, 54 Fed. Reg. 29,597, 29,597-602 (Dep't Commerce July 13, 1989).

[FN78]. There was no assurance that two independent implementations of, say, the FTAM file transfer and access method would interoperate correctly. This is much less of an issue for TCP/IP protocols, where demonstrated interoperability is a prerequisite to standardization. It would be unusual, for instance, for the FTP support in two different TCP/IP implementations to fail to interoperate correctly.

[FN79]. See V. Cerf & K. Mills, RFC 1169: Explaining the Role of GOSIP (Internet Engineering Task Force, Aug. 1990), at

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

http://www.ietf.org/rfc.html.

[FN80]. Approval of Federal Information Processing Standards Publications (FIPS) 146-2, Profiles for Open Systems Internetworking Technologies; and 179- 1, Government Network Management Profile, 60 Fed. Reg. 25,888-02 (Nat'l Inst. of Standards and Tech. May 15, 1995), available at http:// www.itl.nist.gov/fipspubs/fip179-1.htm.

[FN81]. Communications Assistance for Law Enforcement Act, Pub. L. No. 103-414, 108 Stat. 4279 (1994) (codified as amended in scattered sections of 18 U.S.C. and 47 U.S.C.) . For a brief background on CALEA, see FCC, CALEA, at http://www.fcc.gov/calea/ (last reviewed/updated 6/10/04).

[FN82]. Communications Assistance for Law Enforcement Act § 401 (codified as amended at 47 U.S.C. § 1021 (2000)).

[FN83]. In practice, the fund reimburses equipment suppliers. There has been to the author's knowledge only one instance where the fund was used to reimburse a service provider. Service providers incur costs for software upgrades to deploy CALEA, and they incur significant additional deployment costs beyond those associated with hardware and software.

[FN84]. 47 U.S.C. § § 225, 255 (2000).

[FN85]. Id. at § 251.

[FN86]. Cf. I3P Report, supra note 21, at 41 ("Aggressive approaches that more fully use the powers of the federal and state governments are also possible, but the costs and benefits are not well understood and the reasons for a general reluctance to regulate are well known. This statement raises the question of who is responsible for security in this information infrastructure 'commons' and who should pay for it.").

[FN87]. Cf. Hennessy et al., supra note 63, at 103-104 ("it is likely that attempts at such regulation will be fought vigorously, or may fail, because of the likely inability of a regulatory process to keep pace with rapid changes in technology.").

[FN88]. Nat'l Bureau of Standards, A Metric America: A Decision Whose Time Has Come, NBS Special Publication 345, July 1971.

[FN89]. Id. at iii.

[FN90]. Interest in the metric system in the U.S. actually began much earlier. John Quincy Adams considered it in his Report Upon Weights and Measures in 1821. John Quincy Adams, Report on Weights and Measures (1821). Beginning in 1866, a series of laws were enacted that legalized the use of metric weights and measures, and directed the Postmaster General to distribute metric postal scales to all post offices exchanging mail with foreign countries. See Nat'l Bureau of Standards, supra note 88. In fact, the U.S. became the first officially metric country by adopting the metric standards in the Treaty of the Meter to be the nation's "fundamental standards" of weight and mass in 1889. Id. at 14-15.

[FN91]. Metric Conversion Act, Pub. L. No. 94-168, 89 Stat. 1007 (1975) (codified as amended in 15 U.S.C. § 205 (2000)).

[FN92]. Exec. Order No. 12,770, 50 Fed. Reg. 35,801 (July 25, 1991), available at http://ts.nist.gov/ts/htdocs/200/202/pub814.htm#president.

[FN93]. Metric Conversion Act, 89 Stat. 1007.

[FN94]. Id.

[FN95]. Id.

[FN96]. Dr. Gary P. Carver, Nat'l Inst. of Standards & Tech., A Metric America: A Decision Whose Time Has Come - For Real, NISTIR 4858 (1992), available at http://ts.nist.gov/ts/htdocs/200/202/4858.htm (emphasis added). Dr. Carver was then chief of the Metric Program at the National Institutes of Standards and Technology (NIST).

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

[FN97]. Barry M. Leiner et al, Internet Society, A Brief History of the Internet (Dec. 10, 2003), at http://www.isoc.org/internet/history/brief.shtml #Origins. Note that the Advanced Research Projects Agency (ARPA) changed its name to Defense Advanced Research Projects Agency (DARPA) in 1971, then back to ARPA in 1993, and back to DARPA in 1996.

[FN98]. Id.

[FN99]. Id.

[FN100]. See Jason Oxman, The FCC and the Unregulation of the Internet (FCC Office of Plans and Policy, Working Paper No. 31, July 1999), available at http://ftp.fcc.gov/Bureaus/OPP/working_papers/oppwp31.pdf.

[FN101]. Leiner et al., supra note 97.

END OF DOCUMENT

© 2006 Thomson/West. No Claim to Orig. U.S. Govt. Works.

Exhibit H

Modulation

From Wikipedia, the free encyclopedia

For the musical use of "modulation", see ***modulation (music)***.

Modulation is the process of varying a *carrier signal*, typically a sinusoidal signal, in order to use that signal to convey information. The three key parameters of a sinusoid are its amplitude, its phase and its frequency, all of which can be modified in accordance with an information signal to obtain the modulated signal. There are several reasons to modulate a signal before transmission in a medium. These include the ability of different users sharing a medium (multiple access), and making the signal properties physically compatible with the propagation medium. A device that performs modulation is known as a **modulator** and a device that performs the inverse operation of demodulation is known as a **demodulator**. A device that can do both operations is a modem (a contraction of the two terms).

In digital modulation, the changes in the signal are chosen from a fixed list (the **modulation alphabet**) each entry of which conveys a different possible piece of information (a symbol). The alphabet is often conveniently represented on a constellation diagram.

In analog modulation, the change is applied continuously in response to the data signal. The modulation may be applied to various aspects of the signal as the lists below indicate.

Modulation is generally performed to overcome signal transmission issues such as to allow

- Easy (low loss, low dispersion) propagation as electromagnetic waves
- Multiplexing — the transmission of multiple data signals in one frequency band, on different carrier frequencies.
- Smaller, more directional antennas

Carrier signals are usually high frequency electromagnetic waves.

Contents

- 1 Analog modulation techniques
- 2 Digital modulation techniques
- 3 Pulse modulation
- 4 Miscellaneous techniques
- 5 See also
- 6 External links

Analog modulation techniques

- Phase modulation (PM)
- Frequency modulation (FM)
- Amplitude modulation (AM)
  - Single-sideband modulation (SSB, or SSB-AM), very similar to single-sideband suppressed carrier modulation (SSB-SC)
  - Vestigial-sideband modulation (VSB, or VSB-AM)
- Sigma-delta modulation (∑Δ)

Digital modulation techniques

Any form of digital modulation necessarily uses a finite number of distinct signals to represent digital data.

- In the case of PSK, a finite number of phases are used.
- In the case of FSK, a finite number of frequencies are used.
- In the case of ASK, a finite number of amplitudes are used. This is very similar to pulse code modulation

Each of these phases, frequencies or amplitudes are assigned a unique pattern of binary bits. Usually, each phase, frequency or amplitude encodes an equal number of bits. This number of bits comprises the *symbol* that is represented by the particular phase.

These are the general steps used by the modulator to transmit data:

1. Accept incoming digital data;
2. Group the data into symbols;
3. Use these symbols to *set* or *change* the phase, frequency or amplitude of the reference signal appropriately;
4. Pass the modulated signal on for further processing, such as filtering and channel-coding, before transmission.

At the receiver, the demodulator

1. Is passed the de-filtered and de-channel-coded signal;
2. Determines its phase, frequency or amplitude;
3. Maps the phase, frequency or ampltude to its corresponding symbol;
4. Translates the symbol into its individual bits;
5. Passes the resultant bit stream on for further processing such as removal of any error-correcting codes.

As is common to all digital communication systems, the design of both the modulator and demodulator must be done simultaneously. Digital modulation schemes are possible because the transmitter-receiver pair have prior knowledge of how data is encoded and represented in the communications system. In all digital communication systems, both the modulator at the transmitter and the demodulator at the receiver are structured so that they perform inverse operations.

The principal classes of modulation are:

- Phase-shift keying (PSK)
- Frequency-shift keying (FSK) and audio frequency-shift keying (AFSK)
  - Minimum-shift keying (MSK)
  - Gaussian minimum-shift keying (GMSK)
  - Very minimum-shift keying (VMSK)
- Amplitude-shift keying (ASK) and its most common form, on-off keying (OOK)
- Quadrature amplitude modulation (QAM), a combination of PSK and ASK.
- Continuous phase modulation (CPM)
- Trellis coded modulation (TCM) also known as trellis modulation

MSK and GMSK are particular cases of continuous phase modulation (CPM). Indeed, MSK is a particular case of the sub-family of CPM known as continuous phase-frequency-shift keying (CPFSK) which is defined by a rectangular frequency pulse (i.e. a linearly increasing phase pulse) of one symbol-time duration (total response signalling).

Often incorrectly referred to as a modulation scheme, orthogonal frequency division multiplexing (OFDM) usually takes advantage of one of the digital techniques. It is also known as discrete multitone (DMT). When OFDM is used in conjunction with channel coding techniques, it is described as Coded orthogonal frequency division multiplexing (COFDM). OFDM is strictly a channel access method and not a modulation scheme.

Pulse modulation

These are hybrid digital and analogue techniques.

- Pulse-code modulation (PCM)

- Pulse-width modulation (PWM)
- Pulse-amplitude modulation (PAM)
- Pulse-position modulation (PPM)
- Pulse-density modulation (PDM)

Miscellaneous techniques

- The use of on-off keying to transmit Morse code at radio frequencies is known as continuous wave (CW) operation.
- Adaptive modulation
- Wavelet modulation

See also

- Types of radio emissions
- Communications channel
- Channel access methods
- Channel coding
- Line code
- Telecommunication
- Modem
- RF modulator
- Codec

External links

- "Data Encoding Techniques" (http://www.rhyshaden.com/encoding.htm) and "Specifications for Data Encoding" (http://www.wildpackets.com/compendium/FE/FE-Encod.html) discuss the various encoding techniques that have been used with various types of Ethernet.

Retrieved from "http://en.wikipedia.org/wiki/Modulation"

Categories: Disambiguation | Communication theory | Radio modulation modes

---

- This page was last modified 13:10, 20 February 2006.
- All text is available under the terms of the GNU Free Documentation License (see **Copyrights** for details).
  Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc.
- Privacy policy
- About Wikipedia
- Disclaimers

Exhibit I

Attenuation

From Wikipedia, the free encyclopedia

Attenuation is the decrease of the amount, force, magnitude, or value of something. For example,

- In biology, **attenuation** is a mechanism in the regulation of gene expression
- In ecology and geochemistry, **attenuation** is the ability to withhold contaminants in soil and groundwater by various mechanisms like adsorption, dilution, dispersion or biological degradation (biodegradation, bioremediation), causing a decrease in concentration and toxicity compared to the total amount of the contaminant. In environmental engineering and remediation this is often called **natural attenuation**[1] (http://toxics.usgs.gov/definitions/natural_attenuation.html) .
- In wine and beer making, **attenuation** is the measure of thoroughness of fermentation. It is typically given as a percentage number describing how much available sugar has been converted to alcohol during the fermentation process.
- In physical oceanography, light **attenuation** is the decrease in light intensity with depth in the water column due to absorption (by water molecules) and scattering (by suspended particulates).
- In telecommunication, **attenuation** is the decrease in intensity of a signal, beam, or wave as a result of absorption of energy and of scattering out of the path to the detector, but not including the reduction due to geometric spreading.
- In statistics, **attenuation** is another term for regression dilution. See also disattenuation of correlation coefficients.
- In computer industry buzzwords, the word **attenuation** is not clearly defined (http://interconnected.org/home/2005/10/02/attenuation_is) but gaining popularity (http://www.geoplace.com/pressrelease/detail.asp?id=10308) .
- In electronics and audio, **attenuation** is the decrease in amplitude of an electrical signal. **Attenuation** is the opposite of amplification. For example a volume control on an audio system may be referred to as an attenuator.
- In Kant's philosophy, though the word **attenuation** is not mentioned directly, it occurs as **attention** 27 times, and is one of the most important words in Kant's Critique of Pure Reason, his Critical Tables[2] (http://www.bright.net/~jclarke/)

Look up ***Attenuation*** in Wiktionary, the free dictionary

Retrieved from "http://en.wikipedia.org/wiki/Attenuation"

Category: Disambiguation

- This page was last modified 15:31, 20 February 2006.
- All text is available under the terms of the GNU Free Documentation License (see **Copyrights** for details).
  Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc.
- Privacy policy
- About Wikipedia
- Disclaimers

Exhibit J

Decibel

From Wikipedia, the free encyclopedia

The **decibel** (**dB**) is a measure of the ratio between two quantities, and is used in a wide variety of measurements in acoustics, physics and electronics. While originally only used for power and intensity ratios, it has come to be used more generally in engineering. The decibel is widely used as a measure of the loudness of sound. It is a "dimensionless unit" like percent. Decibels are useful because they allow even very large or small ratios to be represented with a conveniently small number. This is achieved by using a logarithm.

Contents

- 1 Definition
  - 1.1 Standards
  - 1.2 Merits
- 2 History of bels and decibels
- 3 Uses
  - 3.1 Acoustics
    - 3.1.1 Frequency weighting
  - 3.2 Electronics
  - 3.3 Optics
  - 3.4 Telecommunications
  - 3.5 Seismology
- 4 Typical abbreviations
  - 4.1 Absolute measurements
    - 4.1.1 Electric power
    - 4.1.2 Electric voltage
    - 4.1.3 Acoustics
    - 4.1.4 Radio power
    - 4.1.5 Note regarding absolute measurements
  - 4.2 Relative measurements
- 5 Reckoning
  - 5.1 Round numbers
  - 5.2 The 4 → 6 energy rule
  - 5.3 The "789" rule
  - 5.4 −3 dB ≈ ½ power
  - 5.5 6 dB per bit
  - 5.6 dB cheat sheet
    - 5.6.1 Commonly used dB values
    - 5.6.2 Other dB values
- 6 See also
- 7 External links
  - 7.1 Converters
- 8 Reference

Definition

An intensity I or power P can be expressed in decibels with the standard equation

$$I_{\mathrm{dB}} = 10\log_{10}\left(\frac{I}{I_0}\right) \quad \text{or} \quad P_{\mathrm{dB}} = 10\log_{10}\left(\frac{P}{P_0}\right) \ ,$$

where I_0 and P_0 are a specified reference intensity and power.

If P_{dB} is 10 dB greater than P_{dB0} then P is ten times P_0. If P_{dB} is 3 dB greater, the power ratio is very close to a factor of two.

For sound intensity, I_0 is typically chosen to be 10^{-12} W/m^2, which is roughly the threshold of hearing. When this choice is made, the units are said to be "dB SIL". For sound power, P_0 is typically chosen to be 10^{-12} W, and the units are then "dB SWL".

In engineering, voltage V or pressure p can be expressed in decibels with the standard equation

$$V_{\mathrm{dB}} = 20\log_{10}\left(\frac{V_1}{V_0}\right) \quad \text{or} \quad p_{\mathrm{dB}} = 20\log_{10}\left(\frac{p_1}{p_0}\right) \ ,$$

where V_0 and p_0 are a specified reference voltage and pressure. Note that in physics, these equations are considered to give *power* in decibels, and it is then incorrect to use them if the electrical or acoustic impedance is not the same at the two points where the voltage or pressure are measured. In this formalism, decibels are always a measure of relative power or intensity, and the value is the same regardless whether power or voltage/pressure measurements are used.

If V_{dB} is 20 dB greater than V_{dB0} then V is ten times V_0. If V_{dB} is 6 dB greater, the voltage ratio is very close to a factor of two.

For sound pressure, p_0 is typically chosen to be 2x10^{-5} N/m^2, or pascals (Pa) which is roughly the threshold of hearing. When this choice is made, the units are said to be "dB SPL".

Standards

The decibel is not an SI unit, although the International Committee for Weights and Measures (BIPM) has recommended its inclusion in the SI system. Following the SI convention, the *d* is lowercase, as it is the SI prefix *deci-*, and the *B* is capitalized, as it is an abbreviation of a name-derived unit, the *bel*, named for Alexander Graham Bell. Written out it becomes *decibel*. This is standard English capitalization.

Merits

The use of decibels has three different merits:

- It is more convenient to add the decibel values of, for instance, two consecutive amplifiers rather than to multiply their amplification factors.
- A very large range of ratios can be expressed with decibel values in a range of moderate size, allowing one to clearly visualize huge changes of some quantity.
- In acoustics, the decibel as a logarithmic measure of ratios fits well to the logarithmic dependence of perceived loudness on sound intensity. In other words, at all levels of loudness, increasing the decibel level by the same amount creates approximately the same increase in perceived loudness — humans perceive the increase from 20 dB to 25 dB as being about the same as the increase from 90 dB to 95 dB, for example. This is known as Stevens' power law.

History of bels and decibels

A **bel** (symbol **B**) is a unit of measure of ratios, such as power levels and voltage levels. It is mostly used in telecommunication, electronics, and acoustics. Invented by engineers of the Bell Telephone Laboratory to quantify the reduction in audio level over a 1 mile length of standard telephone cable, it was originally called the *transmission unit* or *TU*, but was renamed in 1923 or 1924 in honor of the laboratory's founder and telecommunications pioneer Alexander Graham Bell.

The bel was too large for everyday use, so the **decibel (dB)**, equal to 0.1 **bel (B)**, became more commonly used. The bel is still used to represent noise power levels in hard drive specifications.

The neper is a similar unit which uses the natural logarithm. The Richter scale uses numbers expressed in bels as well, though this is implied by definition rather than explicitly stated. In spectrometry and optics, the absorbance unit used to measure optical density is equivalent to −1 B. In astronomy, the apparent magnitude measures the brightness of stars logarithmically, since just as the ear responds logarithmically to acoustic power, the eye responds logarithmically to brightness.

Uses

Acoustics

The **decibel** unit is often used in acoustics to quantify sound levels relative to some 0 dB reference. The reference may be defined as a sound pressure level (SPL), commonly 20 micropascals (20 μPa). To avoid confusion with other decibel measures, the term dB(SPL) is used for this. The reference sound pressure (corresponding to a sound pressure level of 0 dB) can also be defined as the sound pressure at the threshold of human hearing, which is conventionally taken to be 2×10^{-5} newtons per square metre, or 20 micropascals. That is roughly the sound of a mosquito flying 3 m away.

The reason for using the decibel is that the ear is capable of detecting a very large range of sound pressures. The ratio of the sound *pressure* that causes permanent damage from short exposure to the limit that (undamaged) ears can hear is above a million. Because the *power* in a sound wave is proportional to the square of the pressure, the ratio of the maximum power to the minimum power is above one (short scale) trillion. To deal with such a range, logarithmic units are useful: the log of a trillion is 12, so this ratio represents a difference of 120 dB.

Psychologists have found that our perception of loudness is roughly logarithmic — see the Weber-Fechner law. In other words, you have to multiply the sound pressure by the same factor to have the same increase in loudness. This is why the numbers around the volume control dial on a typical audio amplifier are related not to the voltage amplification, but to its logarithm.

Various frequency weightings are used to allow the result of an acoustical measurement to be expressed as a single sound level. The weightings approximate the changes in sensitivity of the ear to different frequencies at different levels. The two most commonly used weightings are the A and C weightings; other examples are the B and Z weightings.

Sound levels above 85 dB are considered harmful, while 120 dB is unsafe and 150 dB causes physical damage to the human body. Windows break at about 163 dB. Jet airplanes cause A-weighted levels of about 133 dB at 33 m, or 100 dB at 170 m. Eardrums rupture at 190 dB to 198 dB. Shock waves and sonic booms cause levels of about 200 dB at 330 m. Sound levels of around 200 dB can cause death to humans and are generated near bomb explosions (e.g., 23 kg of TNT detonated 3 m away). The space shuttle generates levels of around 215 dB (or an A-weighted level of about 175 dB at a distance of 17 m). Even louder are nuclear bombs, earthquakes, tornadoes, hurricanes and volcanoes, all capable of exceeding 240 dB. A more extensive list can be found at makeitlouder.com (http://www.makeitlouder.com/Decibel%20Level%20Chart.txt) .

Some other values:

dB(SPL)	Source (with distance)
250	Inside of tornado; conventional or nuclear bomb explosion at 5 m.

180	Rocket engine at 30 m; blue whale humming at 1 m; Krakatoa explosion at 100 miles (160 km)[1] (http://www.makeitlouder.com/Decibel%20Level%20Chart.txt)
150	Jet engine at 30 m
140	Rifle being fired at 1 m
130	Threshold of pain; train horn at 10 m
120	Rock concert; jet aircraft taking off at 100 m
110	Accelerating motorcycle at 5 m; chainsaw at 1 m
100	Jackhammer at 2 m; inside disco
90	Loud factory, heavy truck at 1 m
80	Vacuum cleaner at 1 m, curbside of busy street
70	Busy traffic at 5 m
60	Office or restaurant inside
50	Quiet restaurant inside
40	Residential area at night
30	Theatre, no talking
10	Human breathing at 3 m
0	Threshold of human hearing (with healthy ears)

Note that the SPL emitted by an object changes with distance from the object. Commonly-quoted measurements of objects like jet engines or jackhammers are meaningless without distance information. The measurement is not of the object's noise, but of the noise at a point in space near that object. For instance, it is intuitively obvious that the noise level of a volcanic eruption will be much higher standing inside the crater than it would be measured from 5 kilometers away.

Measurements of ambient noise do not need a distance, since the noise level will be relatively constant at any point in the area (and are usually only rough approximations anyway).

Measurements that refer to the "threshold of pain" or the threshold at which ear damage occurs are measuring the SPL at a point near the ear itself.

Under controlled conditions, in an acoustical laboratory, the trained healthy human ear is able to discern changes in sound levels of 1 dB, when exposed to steady, single frequency ("pure tone") signals in the mid-frequency range. It is widely accepted that the average healthy ear, however, can barely perceive noise level changes of 3 dB.

On this scale, the normal range of human hearing extends from about 0 dB SPL to about 140 dB SPL. 0 dB SPL is the threshold of hearing in healthy, undamaged human ears at 1 kHz; 0 dB SPL is not an absence of sound, and it is possible for people with exceptionally good hearing to hear sounds at −10 dB SPL. A 3 dB increase in the level of continuous noise doubles the sound power, however experimentation has determined that the response of the human ear results in a perceived doubling of loudness for approximately every 10 dB increase.

Sound pressure levels are applicable to the specific position at which they are measured. The levels change with the distance from the source of the sound; in general, the level decreases as the distance from the source increases. If the distance from the source is unknown, it is difficult to estimate the sound pressure level at the source.

Frequency weighting

Main article: Frequency weighting

Since the human ear is not equally sensitive to all the frequencies of sound within the entire spectrum, noise levels at maximum human sensitivity — middle A and its higher harmonics (between 2,000 and 4,000 hertz) — are factored more heavily into sound descriptions using a process called frequency weighting.

The most widely used frequency weighting is the "A-weighting", which roughly corresponds to the inverse of the 40 dB (at 1 kHz) equal-loudness curve. Using this filter, the sound level meter is less sensitive to very high and very low frequencies. The A weighting parallels the sensitivity of the human ear when it is exposed to normal levels, and frequency weighting C is suitable for use when the ear is exposed to higher sound levels. Other defined frequency weightings, such as B and Z, are rarely used.

Frequency weighted sound levels are still expressed in decibels (with unit symbol dB), although it is common to see the incorrect unit symbols dBA or dB(A) used for A-weighted sound levels.

Electronics

The decibel is used rather than arithmetic ratios or percentages because when certain types of circuits, such as amplifiers and attenuators, are connected in series, expressions of power level in decibels may be arithmetically added and subtracted. It is also common in disciplines such as audio, in which the properties of the signal are best expressed in logarithms due to the response of the ear.

In radio electronics, the decibel is used to describe the ratio between two measurements of electrical power. It can also be combined with a suffix to create an absolute unit of electrical power. For example, it can be combined with "m" for "milliwatt" to produce the "dBm". Zero dBm is one milliwatt, and 1 dBm is one decibel greater than 0 dBm, or about 1.259 mW.

Although decibels were originally used for power ratios, they are commonly used in electronics to describe voltage or current ratios. In a constant resistive load, power is proportional to the square of the voltage or current in the circuit. Therefore, the decibel ratio of two voltages V_1 and V_2 is defined as $20 \log_{10}(V_1/V_2)$, and similarly for current ratios. Thus, for example, a factor of 2.0 in voltage is equivalent to 6.02 dB (not 3.01 dB!). Similarly, a ratio of 10 times gives 20 dB, and one tenth gives −20 dB.

This practice is fully consistent with power-based decibels, provided the circuit resistance remains constant. However, voltage-based decibels are frequently used to express such quantities as the voltage gain of an amplifier, where the two voltages are measured in different circuits which may have very different resistances. For example, a unity-gain buffer amplifier with a high input resistance and a low output resistance may be said to have a "voltage gain of 0 dB", even though it is actually providing a considerable power gain when driving a low-resistance load.

In professional audio, a popular unit is the dBu (see below for all the units). The "u" stands for "unloaded", and was probably chosen to be similar to lowercase "v", as dBv was the older name for the same thing. It was changed to avoid confusion with dBV. This unit (dBu) is an RMS measurement of voltage which uses as its reference 0.775 V_{RMS}. Chosen for historical reasons, it is the voltage level at which you get 1 mW of power in a 600 ohm resistor, which used to be the standard impedance in almost all professional audio circuits.

Since there may be many different bases for a measurement expressed in decibels, a dB value is meaningless unless the reference value (equivalent to 0 dB) is clearly stated. For example, the gain of an antenna system can only be given with respect to a reference antenna (generally a perfect isotropic antenna); if the reference is not stated, the dB gain value is not usable.

Optics

In an optical link, if a known amount of optical power, in dBm (referenced to 1 mW), is launched into a fibre, and the losses, in dB (decibels), of each component (e.g., connectors, splices, and lengths of fibre) are known, the overall link loss may be quickly calculated by simple addition and subtraction of decibel quantities.

Telecommunications

In telecommunications, decibels are commonly used to measure signal-to-noise ratios and other ratio measurements.

Decibels are used to account for the gains and losses of a signal from a transmitter to a receiver through some medium (free space, wave guides, coax, fiber optics, etc.) using a Link Budget.

Seismology

Earthquakes were formerly measured on the Richter scale, which is expressed in bels. (The units in this case are always assumed, rather than explicit.) The more modern moment magnitude scale is designed to produce values comparable to those of the Richter scale.

Typical abbreviations

Absolute measurements

Electric power

dBm *or* dBmW
: dB(1 mW) — power measurement relative to 1 milliwatt.

dBW
: dB(1 W) — same as dBm, with reference level of 1 watt.

Electric voltage

dBu *or* dBv
: dB(0.775 V) — (usually RMS) voltage amplitude referenced to 0.775 volt. Although dBu can be used with any impedance, dBu = dBm when the load is 600Ω. dBu is preferable, since dBv is easily confused with dBV. The "u" comes from "unloaded".

dBV
: dB(1 V) — (usually RMS) voltage amplitude of a signal in a wire, relative to 1 volt, not related to any impedance.

Acoustics

dB(SPL)
: dB(Sound Pressure Level) — relative to 20 micropascals (μPa) = 2×10^{-5} Pa, the quietest sound a human can hear. This is roughly the sound of a mosquito flying 3 metres away. This is often abbreviated to just "dB", which gives some the erroneous notion that a dB is an absolute unit by itself.

Radio power

dBm
: dB(mW) — power relative to 1 milliwatt.

dBμ *or* dBu
: dB(μV/m) — electric field strength relative to 1 microvolt per metre.

dBf
: dB(fW) — power relative to 1 femtowatt.

dBW
: dB(W) — power relative to 1 watt.

dBk
: dB(kW) — power relative to 1 kilowatt.

Note regarding absolute measurements

The term "measurement relative to" means so many dB greater, or smaller, than the quantity specified.

Examples:

- 3 dBm means 3 dB greater than 1 mW.
- −6 dBm means 6 dB less than 1 mW.
- 0 dBm means no change from 1 mW, in other words 0 dBm *is* 1 mW.

Relative measurements

dB(A), dB(B), and dB(C) weighting
: These symbols are often used to denote the use of different frequency weightings, used to approximate the human ear's response to sound, although the measurement is still in dB (SPL). Other variations that may be seen are dB_A or dBA. According to ANSI standards, the preferred usage is to write L_A = x dB, as dBA implies a reference to an "A" unit, not an A-weighting. They are still used commonly as a shorthand for A-weighted measurements, however.

dBd
: dB(dipole) — the forward gain of an antenna compared to a half-wave dipole antenna.

dBi
: dB(isotropic) — the forward gain of an antenna compared to an idealized isotropic antenna.

dBFS *or* dBfs
: dB(full scale) — the amplitude of a signal (usually audio) compared to the maximum which a device can handle before clipping occurs. In digital systems, 0 dBFS would equal the highest level (number) the processor is capable of representing. (Measured values are negative, since they are less than the maximum.)

dBr
: dB(relative) — simply a relative difference to something else, which is made apparent in context. The difference of a filter's response to nominal levels, for instance.

dBrn
: dB above reference noise.

dBC
: dB relative to carrier — in fiberoptic telecommunications, this indicates the relative levels of noise or sideband peak power, compared to the optical carrier power.

Reckoning

Decibels are handy for mental calculation, because adding them is easier than multiplying ratios. First, however, one has to be able to convert easily between ratios and decibels. The most obvious way is to memorize the logs of small primes, but there are a few other tricks that can help.

Round numbers

The values of coins and banknotes are round numbers. The rules are:

1. One is a round number
2. Twice a round number is a round number: 2, 4, 8, 16, 32, 64
3. Ten times a round number is a round number: 10, 100
4. Half a round number is a round number: 50, 25, 12.5, 6.25
5. The tenth of a round number is a round number: 5, 2.5, 1.25, 1.6, 3.2, 6.4

Now 6.25 and 6.4 are approximately equal to 6.3, so we don't care. Thus the round numbers between 1 and 10 are these:

```
Ratio  1    1.25 1.6  2    2.5  3.2  4    5    6.3  8    10
dB     0    1    2    3    4    5    6    7    8    9    10
```

This useful approximate table of logarithms is easily reconstructed or memorized.

The 4 → 6 energy rule

To one decimal place of precision, 4.x is 6.x in dB (energy).

Examples:

- 4.0 → 6.0 dB
- 4.3 → 6.3 dB
- 4.7 → 6.7 dB

The "789" rule

To one decimal place of precision, x → (½ x + 5.0 dB) for 7.0 ≤ x ≤ 10.

Examples:

- 7.0 → ½ 7.0 + 5.0 dB = 3.5 + 5.0 dB = 8.5 dB
- 7.5 → ½ 7.5 + 5.0 dB = 3.75 + 5.0 dB = 8.75 dB
- 8.2 → ½ 8.2 + 5.0 dB = 4.1 + 5.0 dB = 9.1 dB
- 9.9 → ½ 9.9 + 5.0 dB = 4.95 + 5.0 dB = 9.95 dB
- 10.0 → ½ 10.0 + 5.0 dB = 5.0 + 5.0 dB = 10 dB

−3 dB ≈ ½ power

A level difference of ±3 dB is roughly double/half power (equal to a ratio of 1.995). That is why it is commonly used as a marking on sound equipment and the like.

Another common sequence is 1, 2, 5, 10, 20, 50 ... . These preferred numbers are very close to being equally spaced in terms of their logarithms. The actual values would be 1, 2.15, 4.64, 10 ... .

The conversion for decibels is often simplified to: "+3 dB means two times the power and 1.414 times the voltage", and "+6 dB means four times the power and two times the voltage ".

While this is accurate for many situations, it is not exact. As stated above, decibels are defined so that +10 dB means "ten times the power". From this, we calculate that +3 dB actually multiplies the power by $10^{3/10}$. This is a power ratio of 1.9953 or about 0.25% different from the "times 2" power ratio that is sometimes assumed. A level difference of +6 dB is 3.9811, about 0.5% different from 4.

To contrive a more serious example, consider converting a large decibel figure into its linear ratio, for example 120 dB. The power ratio is correctly calculated as a ratio of 10^{12} or one trillion. But if we use the assumption that 3 dB means "times 2", we would calculate a power ratio of $2^{120/3} = 2^{40} = 1.0995 \times 10^{12}$, giving a 10% error.

6 dB per bit

In digital audio, each bit offered by the system doubles the (voltage) resolution, corresponding to a 6 dB ratio. For instance, a 16-bit (linear) audio format offers an approximate theoretical maximum of (16 × 6) = 96 dB, meaning that the maximum signal (see *0 dBFS*, above) is 96 dB above the quantization noise.

dB cheat sheet

As is clear from the above description, the dB level is a logarithmic way of expressing power ratios. The following tables are cheat-sheets that provide values for various dB levels.

Commonly used dB values

dB level	Ratio
−30 dB	1/1000
−20 dB	1/100
−10 dB	1/10
−3 dB	0.5 (approx.)
3 dB	2 (approx.)
10 dB	10
20 dB	100
30 dB	1000

Other dB values

dB level	Ratio
−9 dB	1/8 (approx.)
−6 dB	1/4 (approx.)
−1 dB	0.8 (approx.)
1 dB	1.25 (approx.)
6 dB	4 (approx.)
9 dB	8 (approx.)

See also

- Equal-loudness contour
- ITU-R 468 noise weighting
- Noise (environmental)
- Signal noise
- Sound pressure level
- Weighting filter—discussion of **dBA**
- Decibel magazine

External links

- What is a decibel? (http://www.phys.unsw.edu.au/~jw/dB.html)
- Lindos Electronics Audio Articles (http://www.lindos.co.uk/cgi-bin/FlexiData.cgi?SOURCE=Articles)
- Description of some abbreviations (http://www.sizes.com/units/decibel.htm)
- Noise Control and Hearing Conservation (http://www.uoguelph.ca/HR/ehs/policies/10-01.pdf)
- Noise Measurement OSHA 1 (http://www.osha.gov/dts/osta/otm/otm_iii/otm_iii_5.html)
- Noise Measurement OSHA 2 (http://www.environmental-center.com/articles/article138/article138.htm)
- Understanding dB (http://www.jimprice.com/prosound/db.htm)
- Rane Professional Audio Reference entry for "decibel" (http://www.rane.com/par-d.html#decibel)
- Hyperphysics description of decibels (http://hyperphysics.phy-astr.gsu.edu/hbase/sound/db.html#c1)
- Decibel Magazine (http://decibelmagazine.com)

Converters

- V_{peak}, V_{RMS}, Power, dBm, dBu, dBV converter (http://www.analog.com/Analog_Root/static/techSupport/designTools/interactiveTools/dbconvert/dbconvert.html)
- Conversion: dBu to volts, dBV to volts, and volts to dBu, and dBV (http://www.sengpielaudio.com/calculator-db-volt.htm)
- Conversion of sound level units: dBSPL or dBA to sound pressure p and sound intensity J (http://www.sengpielaudio.com/calculator-soundlevel.htm)
- Conversion: Voltage V to dB, dBu, dBV, and dBm (http://www.sengpielaudio.com/calculator-volt.htm)
- Only Power: dBm to mW conversion (http://www.moonblinkwifi.com/dbm_to_watt_conversion.cfm)

Reference

- Martin, W. H., "DeciBel – The New Name for the Transmission Unit", *Bell System Technical Journal*, January 1929.

Retrieved from "http://en.wikipedia.org/wiki/Decibel"

Categories: Units of measure | Sound | Acoustics

- This page was last modified 06:05, 5 March 2006.
- All text is available under the terms of the GNU Free Documentation License (see **Copyrights** for details).
  Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc.
- Privacy policy
- About Wikipedia
- Disclaimers

Exhibit K

Value-Added Module System

LGX Compatible

ADC's Value-Added Module (VAM) System can now be incorporated into installed LGX fiber frames. ADC's LGX-compatible products encourage lineup consistency while improving fiber cable management. ADC's LGX-compatible VAMs allow service providers to incorporate optical splitters for circuit monitoring and video distribution. VAMs can also be configured with wavelength division multiplexing capabilities to increase transmission capacity over existing fiber lines. Various input and output interface options are available.

Features:

- **Enclosed plug-in modules**
  Optical components are protected from physical and environmental damage
- **Removable straight retainers**
  Maintain minimum bend radius of fiber patch cords and allow single fiber access
- **Flexible platform**
  Modules can be created for new applications quickly and easily to meet customer requirements
- **Monitor and/or test**
  Enables providers to troubleshoot networks without forcing disruption of service
- **Custom configurations**
  Custom splitter configurations available upon request

Value-Added Module System

LGX Compatible

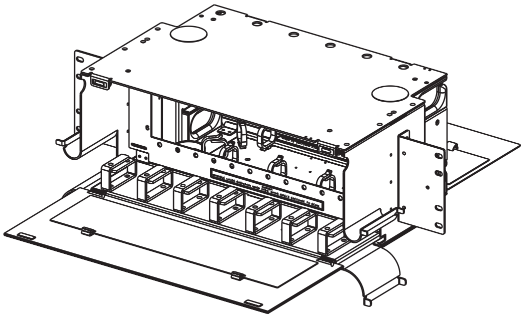
LGX Chassis

The 7" (17.78 cm) LGX VAM chassis is designed to fit into any open chassis location within new or existing LGX-compatible fiber frames. It can accommodate a maximum of 12 plug-in modules, 12 bulkhead plates, 12 blank plates, or any combination thereof. The mounting slots are oriented vertically.

The 3.5" (8.89 cm) LGX VAM chassis is designed to fit into any 19" (48.26 cm) or 23" (58.42 cm) environment. It can accommodate a maximum of four plug-in modules, four bulkhead plates, four blank plates or any combination thereof. The mounting slots are oriented horizontally to allow standard LGX plug-in modules to be used while maximizing density and minimizing rack space allotment.

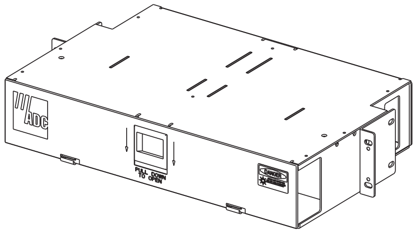
Ordering Information

Description	Dimensions (HxWxD)	Catalog Number
7" unloaded chassis; VAMs mount vertically; accommodates up to 12 modules	7" x 19/23" x 11" (17.78 x 48.26/58.42 x 27.94 cm	FVM-VLM19X700-W



Ordering Information

Description	Dimensions (HxWxD)	Catalog Number
3.5" unloaded chassis; VAMs mount horizontally; accommodates up to 4 modules	3.50" x 18.76" x 11" (8.89 x 47.65 x 27.94 cm)	FVM-LGX19X350-W



Value-Added Module System
LGX Compatible

Video Splitter Module Ordering Information

Catalog Number

VLM- _ __ __ ____ _ __ __ _

Module Option

2	Input front, output front
5	Input rear, output front

Input Connector/Adapter
Singlemode

2U	FC ultra polish
2A	FC angled polish
4U	ST® ultra polish
7U	SC ultra polish
7A	SC angled polish
8A	E-2000 angled polish
LA	LX.5® angled polish
00	Bare fiber

or

Multimode**

5F	ST® 50/125
5U	ST® 62.5/125
9A	SC 50/125
9B	SC 62.5/125
LM	LX.5® 62.5/125
LN	LX.5® 50/125
2M	FC 62.5/125
2N	FC 50/125
00	Bare fiber

Output Connector/Adapter
Singlemode

2U	FC ultra polish
2A	FC angled polish
4U	ST® ultra polish
7U	SC ultra polish
7A	SC angled polish
8A	E-2000 angled polish
LA	LX.5® angled polish
00	Bare fiber

or

Multimode**

5F	ST® 50/125
5U	ST® 62.5/125
9A	SC 50/125
9B	SC 62.5/125
LM	LX.5® 62.5/125
LN	LX.5® 50/125
2M	FC 62.5/125
2N	FC 50/125
00	Bare fiber

Split Ratio

2000	1x2	50/50
2100	1x2	55/45
2200	1x2	60/40
2300	1x2	65/35
2350	1x2	66.6/33.3
2400	1x2	70/30
2500	1x2	75/25
2600	1x2	80/20
2700	1x2	85/15
2800	1x2	90/10
2900	1x2	95/5
2950	1x2	99/1
2030	1x2	50/50
3900	1x3	33/33/33
4900	1x4	25/25/25/25
5300	1x5	20/20/20/20/20
6000	1x6	16.6/16.6/...16.6
7000	1x7	14.3/14.3/...14.3
8000	1x8	12.5/12.5/...12.5
9000	1x9	11.1/11.1/...11.1
A110	1x10	10/10/10/...10
B111	1x11	9.1/9.1/...9.1
C112	1x12	8.33/8.33...8.33
D113	1x13	7.7/7.7/...7.7
E114	1x14	7.14/7.14...7.14
F115	1x15	6.66/6.66/...6.66
G116	1x16	6.25/6.25/...6.25
M132	1x17	3.125/3.125/...3.125

Splitter Type

LEAVE BLANK	Wideband 1310 and 1550 nm
G	1310 nm wavelength flattened
J	1550 nm wavelength flattened

Output Pigtail

Length† (in meters)
Example:
00 = No pigtail
01 = 1 meter
10 = 10 meters

Input Pigtail

Length† (in meters)
Example:
00 = No pigtail
01 = 1 meter
10 = 10 meters

† Leave 00 for adapter option

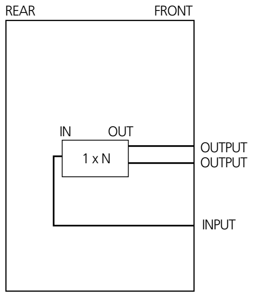
Number of Splitters

A	Single splitter
B	Dual splitter
C	Triple splitter

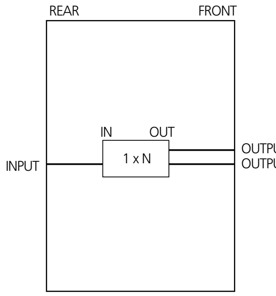
Other split ratios are available. Call ADC Technical Assistance Center at 1-800-366-3891, Ext. 73475 or visit our website at www.adc.com.

**Contact ADC for multimode specifications and available split ratios

2 Input Front, Output Front



5 Input Rear, Output Front



Value-Added Module System
LGX Compatible

Singlemode Wideband Optical Splitter Specifications

Other splitter configurations are available. Please call ADC Technical Assistance Center, 1-800-366-3891, ext. 73475, or visit our website at www.adc.com.

Splitter Number	Splitter Type	Splitter Ratio	Maximum Insertion Loss*	Typical Insertion Loss*
2000	1x2	50/50	3.7/3.7	3.1/3.1
2100	1x2	55/45	3.2/4.1	2.7/3.6
2200	1x2	60/40	2.7/4.7	2.3/4.1
2300	1x2	65/35	2.3/5.3	2.0/4.7
2350	1x2	66.6./33.3	2.2/5.5	1.9/5.0
2400	1x2	70/30	2.0/6.0	1.7/5.4
2500	1x2	75/25	1.6/6.8	1.4/6.2
2600	1x2	80/20	1.3/7.8	1.1/7.1
2700	1x2	85/15	1.0/9.2	.8/8.4
2800	1x2	90/10	.8/11.2	.6/10.2
2900	1x2	95/5	.5/14.4	.4/13.2
2950	1x2	99/1	0.3/22.5	NA/NA
2030	2x2	50/50	3.7/3.7	3.1/3.1
3900	1x3	33/33/33	5.9/5.9/5.9	5.0/5.0/5.0
4900	1x4	25/25/25/25	7.4/7.4/7.4/7.4	6.2/6.2/6.2/6.2
5300	1x5	20/20/20/20/20	8.6/8.6/8.6/8.6/8.6	7.3/7.3/7.3/7.3/7.3
6000	1x6	16.6/16.6/ . . . 16.6	9.8/9.8/9.8/ . . ./9.8	8.1/8.1/8.1 . . . /8.1
7000	1x7	14.3/14.3/ . . . /14.3	10.6/10.6/ . . . /10.6	8.9/8.9/ . . . /8.9
8000	1x8	12.5/12.5/ . . .12.5	11.3/11.3/ . . . /11.3	9.5/9.5/ . . ./9.5
9000	1x9	11.1/11.1/ . . . /11.1	11.5/11.5/ . . . /11.5	9.9/9.9/ . . . /9.9
A110	1x10	10/10/10 . . . /10	12.1/12.1/ . . . /12.1	10.4/10.4/ . . . /10.4
B111	1x11	9.1/9.1/ . . . /9.1	13.1/13.1/ . . . /13.1	10.9/10.9/ . . . /10.9
C112	1x12	8.3/8.3/ . . . /8.3	13.3/13.3/ . . . /13.3	11.3/11.3/ . . . /11.3
D113	1x13	7.7/7.7/ . . . /7.7	13.8/13.8/ . . . /13.8	11.6/11.6/ . . . /11.6
E114	1x14	7.14/7.14/ . . . /7.14	13.8/13.8/ . . . /13.8	11.8/11.8/ . . . /11.8
F115	1x15	6.66/6.66/ . . . /6.66	14.5/14.5/ . . . /14.5	12.4/12.4/ . . . /12.4
G116	1x16	6.25/6.25/ . . . 6.25	14.9/14.9/ . . . 14.9	12.6/12.6/ . . . 12.6
M132	1x32	3.13/3.13/ . . . 3.13	18.5/18.5/ . . . 18/5	15.8/15.8/ . . . 15.8

* Does not include connector loss. See table on the following page or call ADC Technical Assistance Center, 1-800-366-3891, ext. 73475.

Maximum Insertion Loss
Maximum insertion loss is the upper limit of insertion loss for the coupler and applies over the entire wavelength range specified in the bandpass.

Typical Insertion Loss
Typical insertion loss is the expected insertion loss value for the coupler measured at the specified center wavelength (i.e., 1310 nm and/or 1550 nm).

Value-Added Module System
LGX Compatible

Singlemode Optical Splitter Specifications

MATERIAL

Fiber:	Corning SMF-28
Jacket:	0.900 mm hytrel tube
Housing:	Stainless steel tube

OPTICAL

Bandpass	
Wavelength 1:	1260/1360 nm
Wavelength 2:	1430/1580 nm
Directivity:	55 dB minimum
Reflectance:	< -55 dB
Polarization Stability:	.5 dB maximum

MECHANICAL

Cable Retention:	.45 kg (1.0 lb.) maximum for .25 mm and .900 mm fiber, 1 kg (2.2 lbs.) maximum for cabled fiber assemblies
Flex:	300 cycles; .45 kg (1.0 lb.) maximum load (fiber cable only)
Twist:	100 cycles; .45 kg (1.0 lb.) maximum load (fiber cable only)
Vibration:	10-55 Hz per FOTP II condition I
Impact:	8 drops from 1.5 meters
Flammability:	All materials used in fabrication of coupler housing meet UL 94, Level V-I requirements

ENVIRONMENTAL

Humidity:	336 hours, 60°C at 95% relative humidity
Thermal Aging:	336 hours, 85°C
Temperature Cycling:	336 hours, -40°C to 85°C

Singlemode Optical Connector Specifications

	INSERTION LOSS (dB)	RETURN LOSS (dB)
Ultra Polish PC Connectors:	0.3 dB maximum 0.15 dB typical	57 dB minimum
Angled Polish Connectors:	0.5 dB maximum 0.2 dB typical	65 dB minimum

Value-Added Module System
LGX Compatible

WDM Module Ordering Information

Catalog Number

VLM-W _ __ __ _ _ _ _ __ __

Configuration

2	MUX front, DEMUX front
5	MUX rear, DEMUX front

Mux Port Connector/Adapter

2U	FC ultra polish
2A	FC angled polish
4U	ST® ultra polish
7U	SC ultra polish
7A	SC angled polish
8A	E-2000 angled polish
LA	LX.5® angled polish
00	Bare fiber

Demux Port Connector/Adapter

2U	FC ultra polish
2A	FC angled polish
4U	ST® ultra polish
7U	SC ultra polish
7A	SC angled polish
8A	E-2000 angled polish
LA	LX.5® angled polish
00	Bare fiber

Module Configuration

A	Single
B	Dual
C	Triple

WDM Type (λ_1, λ_2)

A	Undirectional multiplexer
C	Undirectional demultiplexer
E	Bidirectional λ_1 Tx/λ_2 Rx
G	Bidirectional λ_1 Rx/λ_2 Tx
J	45 dB isolation, pass 1310
K	45 dB isolation, pass 1550
L	(Dual only) unidirectional multiplexer, undirectional demultiplexer*
M	(Dual only) bidirectional λ_1 Tx/λ_2 Rx, bidirectional λ_2 Tx/λ_1 Rx

* WDM Type "L" includes (1) unidirectional multiplexer (standard isolation) and (1) unidirectional demultiplexer (isolation specified by next character)

Demux Port Pigtail Length

Length† (in meters)
Example:
00 = No pigtail
01 = 1 meter
10 = 10 meters

Mux Port Pigtail Length

Length† (in meters)
Example:
00 = No pigtail
01 = 1 meter
10 = 10 meters

† Leave 00 for adapter option

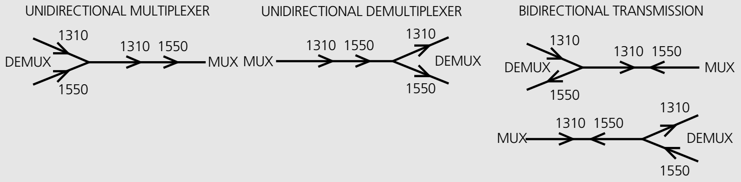
Wavelength

A	1310 (λ_1)/1550(λ_2)
B	1533 (λ_1)/1557(λ_2)
E	1550 (λ_1)/1625(λ_2)
F	1310 (λ_1)/1625(λ_2)

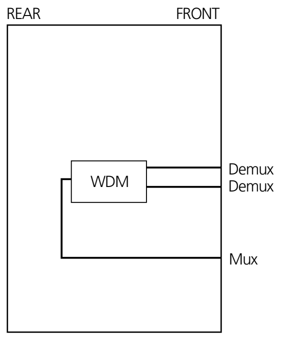
Isolation

1	Standard
2	High
3	Very high

In the ordering charts, the abbreviation "mux" references the multiplexed side of the WDM, the side where two optical signals coexist on one fiber. The abbreviation "demux" references the demultiplexed side of the WDM, the side where each signal appears on its own fiber. Both unidirectional and bidirectional WDMs are available as shown below.

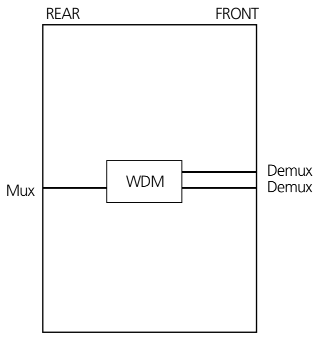

UNIDIRECTIONAL MULTIPLEXER
1310
DEMUX
1550
1310 1550
MUX
UNIDIRECTIONAL DEMULTIPLEXER
MUX
1310 1550
1310
DEMUX
1550
BIDIRECTIONAL TRANSMISSION
1310
DEMUX
1550
1310 1550
MUX
MUX
1310 1550
1310
DEMUX
1550

2 Mux Front, Demux Front



REAR
FRONT
WDM
Demux
Demux
Mux

5 Mux Rear, Demux Front



REAR
FRONT
Mux
WDM
Demux
Demux

Value-Added Module System

LGX Compatible

Singlemode WDM Specifications

MATERIAL

Fiber:	Corning SMF-28
Jacket:	0.900 mm

OPTICAL

Uniformity:	1.0 Log2N (where N = number of channels)
Polarization Stability:	0.5 dB maximum
Bandpass:	(see table below)
Insertion Loss:	(see table below)
Reflectance:	< -55 dB
Isolation	
Near End:	55 dB
Far End:	(see table below)

MECHANICAL

Cable Retention:	.45 kg (1.0 lb.) maximum for .250 mm and .900 mm fiber; 1.0 kg (2.2 lbs.) maximum for cabled fiber assemblies
Flex:	300 cycles; .45 kg (1.0 lb.) maximum load (fiber cable only)
Twist:	100 cycles; .45 kg (1.0 lb.) maximum load (fiber cable only)
Vibration:	10-55 Hz per FOTP II condition I
Impact:	8 drops from 1.5 meters
Flammability:	All materials used in fabrication of the coupler housing meet UL 94, Level V-I requirements

ENVIRONMENTAL

Humidity Resistance:	336 hours, 60°C at 95% relative humidity
Thermal Aging:	336 hours, 85°C
Temperature Cycling:	336 hours, -40°C to 85°C

1310/1550 nm	PASSBAND (nm)	MAXIMUM INSERTION LOSS* (dB)	MINIMUM ISOLATION (dB)
Multiplexer/Demultiplexer:	1310 ± 20 nm and 1550 ± 20 nm	0.3	15
High Isolation Demultiplexer:	1310 ± 20 nm and 1550 ± 20 nm	0.7	30
Very High Isolation Demultiplexer:	1310 ± 20 nm and 1550 ± 20 nm	1.0	45
Wavelength Filters:	1310 ± 20 nm and 1550 ± 20 nm	0.3	15

1533/1557 nm	PASSBAND (nm)	MAXIMUM INSERTION LOSS* (dB)	MINIMUM ISOLATION (dB)
Unidirectional Multiplexer:	1533 ±3 and 1557 ±3 nm	0.9	N/A
High Isolation Undirectional Multiplexer or Bidirectional Multiplexer/Demultiplexer:	1533 ±3 and 1557 ±3 nm	1.8	20
Very High Isolation Undirectional Multiplexer or Bidirectional Multiplexer/Demultiplexer:	1533 ±3 and 1557 ±3 nm	2.1	35
Filter Passing 1533 nm:	1533 ±3 nm	0.9	13
Filter Passing 1557 nm:	1557 ±3 nm	0.9	13

* Maximum Insertion Loss does not include connector loss. See pages 4-5 for connector specifications and definitions.

Value-Added Module System

LGX Compatible

Monitor Module Ordering Information

Catalog Number

VLM-M _ _ _ _ _ _ _ _ _

Input Connector Type Singlemode

2U	FC ultra polish
2A	FC angled polish
4U	ST® ultra polish
7U	SC ultra polish
7A	SC angled polish
8A	E-2000 angled polish
LA	LX.5® angled polish
00	Bare fiber

or

Multimode**

5F	ST® 50/125
5U	ST® 62.5/125
9A	SC 50/125
9B	SC 62.5/125
LM	LX.5® 62.5/125
LN	LX.5® 50/125
2M	FC 62.5/125
2N	FC 50/125
00	Bare fiber

Output Connector Type Singlemode

2U	FC ultra polish
2A	FC angled polish
4U	ST® ultra polish
7U	SC ultra polish
7A	SC angled polish
8A	E-2000 angled polish
LA	LX.5® angled polish
00	Bare fiber

or

Multimode**

5F	ST® 50/125
5U	ST® 62.5/125
9A	SC 50/125
9B	SC 62.5/125
LM	LX.5® 62.5/125
LN	LX.5® 50/125
2M	FC 62.5/125
2N	FC 50/125
00	Bare fiber

Number of Circuits per VAM

1	1 circuit
2	2 circuits
3	3 circuits

Split Ratio

A	90/10
B	95/05
C	50/50
E	80/20
F	75/25
H	70/30
J	60/40
K	99/01

Monitor Connector Type Singlemode

2U	FC ultra polish
2A	FC angled polish
4U	ST® ultra polish
7U	SC ultra polish
7A	SC angled polish
8A	E-2000 angled polish
LA	LX.5® angled polish
00	Bare fiber

or

Multimode**

5F	ST® 50/125
5U	ST® 62.5/125
9A	SC 50/125
9B	SC 62.5/125
LM	LX.5® 62.5/125
LN	LX.5® 50/125
2M	FC 62.5/125
2N	FC 50/125
00	Bare fiber

Monitor Pigtail Length

Output Pigtail Length

Input Pigtail Length

Length† (in meters)
Example:
00 = No pigtail
01 = 1 meter
10 = 10 meters

† Leave 000 for adapter option

Module Style

B	Cross-connect customer premises
D	Dual monitor cross-connect
G	Dual monitor cross-connect
J	Dual monitor
P	Customer premises
R	Single monitor cross-connect

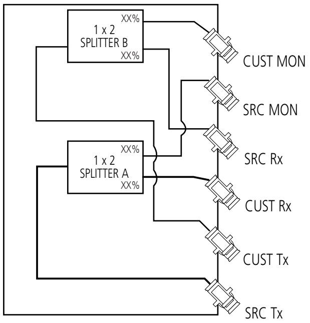
* See detailed splitter and connector specifications on pages 4-5.

Contact ADC for multimode specifications and available split ratios at 1-800-366-3891, Ext. 73475 or visit our website at www.adc.com.

Value-Added Module System

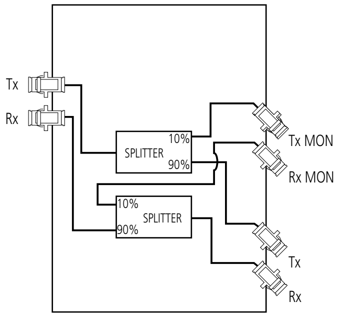
LGX Compatible

B Customer Premises Monitor Module



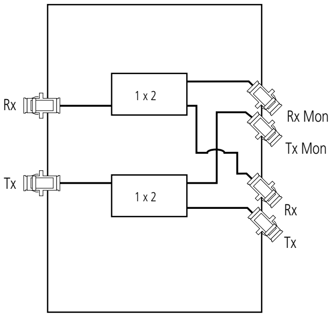
XX%
1 x 2
SPLITTER B
XX%
XX%
1 x 2
SPLITTER A
XX%
CUST MON
SRC MON
SRC Rx
CUST Rx
CUST Tx
SRC Tx

D Dual Monitor Cross-Connect



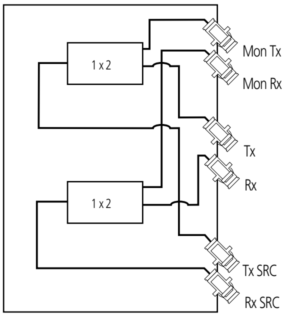
Tx
Rx
SPLITTER
10%
90%
10%
SPLITTER
90%
Tx MON
Rx MON
Tx
Rx

G Dual Monitor Cross-Connect



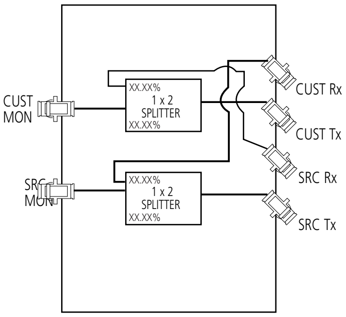
Rx
Tx
1 x 2
1 x 2
Rx Mon
Tx Mon
Rx
Tx

J Dual Monitor Cross-Connect



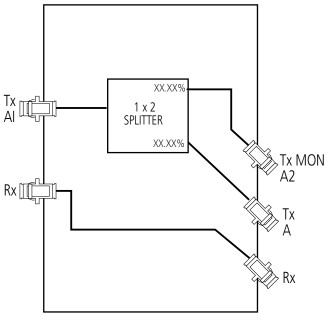
1 x 2
1 x 2
Mon Tx
Mon Rx
Tx
Rx
Tx SRC
Rx SRC

P Customer Premises Monitor Module



CUST
MON
SRC
MON
XX.XX%
1 x 2
SPLITTER
XX.XX%
XX.XX%
1 x 2
SPLITTER
XX.XX%
CUST Rx
CUST Tx
SRC Rx
SRC Tx

R Single Monitor Cross-Connect



Tx
AI
Rx
XX.XX%
1 x 2
SPLITTER
XX.XX%
Tx MON
A2
Tx
A
Rx

Web Site: www.adc.com
From North America, Call Toll Free: 1-800-366-3891, Ext. 73475 • Outside of North America: +1-952-938-8080 Fax: +1-952-946-3292
For a complete listing of ADC's global sales office locations, please refer to our web site.

ADC Telecommunications, Inc., P.O. Box 1101, Minneapolis, Minnesota USA 55440-1101
Specifications published here are current as of the date of publication of this document. Because we are continuously improving our products, ADC reserves the right to change specifications without prior notice. At any time, you may verify product specifications by contacting our headquarters office in Minneapolis. ADC Telecommunications, Inc. views its patent portfolio as an important corporate asset and vigorously enforces its patents. Products or features contained herein may be covered by one or more U.S. or foreign patents. An Equal Opportunity Employer

CERTIFIED ISO9001 CERTIFIED

891-LGX 12/01 Original © 2001   ADC Telecommunications, Inc. All Rights Reserved